



**Spring 2023 NSCAS Growth
ELA, Mathematics, and Science Technical Report**

Table of Contents

Executive Summary	11
Section 1: Introduction	14
1.1. NSCAS Overview	14
1.2. Background.....	14
1.3. Schedule of Major Events.....	15
1.4. Building a Validity Argument.....	16
1.5. Intended Purposes and Uses of Test Results	17
1.6. Theory of Action.....	18
Section 2: Test Design and Development.....	20
2.1. Test Designs.....	20
2.2. Academic Content Standards	24
2.3. Blueprints.....	24
2.4. Item Types.....	24
2.5. Depth of Knowledge (DOK)	25
2.6. ALD Development.....	26
2.6.1. Policy ALDs	26
2.6.2. Range ALDs	26
2.6.2.1. ELA and Mathematics.....	27
2.6.2.2. Science	30
2.6.3. Reporting ALDs	31
2.7. Item Development.....	32
2.7.1. Item Specifications.....	32
2.7.2. Science.....	32
2.7.3. Item Retirement	34
2.8. Content Alignment	34
2.8.1. Alignment and Adaptive Testing	34
2.8.2. 2023 ELA Alignment Study.....	35
2.9. Universal Design.....	36
2.10. Sensitivity and Fairness.....	36
2.11. Test Construction (ELA and Mathematics)	37
2.11.1. Fixed-Forms.....	37
2.11.2. MAP Growth Item Selection.....	37
2.12. Data Review	37
Section 3: Test Administration and Security	40
3.1. User Roles and Responsibilities	41
3.2. Administration Training.....	41
3.3. Item Type Samplers.....	42
3.4. Accommodations and Accessibility Features	42
3.5. User Acceptance Testing (UAT)	45
3.6. Student Participation.....	45
3.6.1. Paper-Pencil Participation Criteria.....	46
3.6.2. Participation of English Language Learners (ELLs)	46

3.6.3. Participation of Recently Arrived Limited English Proficient Students	47
3.7. Test Security	47
3.7.1. Test Security	47
3.7.1.1. Physical Warehouse Security	47
3.7.1.2. Secure Destruction of Test Materials	48
3.7.1.3. Shipping Security	48
3.7.1.4. Electronic Security of Test Materials and Data	48
3.8. Partner Support	48
Section 4: Scoring and Reporting	50
4.1. Scoring Rules	50
4.2. Score Reporting Methods	51
4.3. Report Summary	52
4.3.1. Report Verification	54
Section 5: Adaptive Engine	56
5.1. Overview	56
5.2. Engine Simulations and Evaluation	57
5.2.1. Evaluation Criteria	58
5.2.2. Blueprint Constraint Accuracy	59
5.2.3. Item Exposure Rates	66
5.2.4. Score Precision and Reliability	71
5.3. Engine Simulations: Science Field Test	85
Section 6: Psychometric Analyses	93
6.1. Number of Students Included in the Analyses	93
6.2. Classical Item Analyses	94
6.2.1. Item Difficulty (P Value)	94
6.2.2. Item Discrimination (Item-Total Correlation)	97
6.2.3. Item Suppression	100
6.3. Differential Item Functioning (DIF)	102
6.3.1. Logistic Regression (LR) DIF Method	102
6.3.2. Mantel-Haenszel (MH) DIF Methods	104
6.3.3. DIF Results	106
6.4. IRT Calibration	109
6.4.1. Summary of IRT Item Statistics	110
6.5. Scaling	111
Section 7: Standard Setting	114
7.1. ELA and Mathematics	114
7.1.1. Methodology	114
7.1.2. Meeting Process	114
7.1.3. Articulation	116
7.2. Science	116
7.2.1. Methodology	116
7.2.2. Meeting Process	117
7.3. Final Results	117
Section 8: Test Results	119

8.1. Demographics and Accommodations	119
8.2. Administration Mode (Online vs. Paper-Pencil)	125
8.3. Testing Time	126
8.4. Achievement Level Distributions.....	129
8.5. Descriptive Statistics of Scale Scores	130
8.6. Reporting Category Correlations	132
8.7. Correlations with MAP Growth.....	135
Section 9: Reliability.....	136
9.1. Marginal Reliability.....	136
9.2. Conditional Standard Error of Measurement (CSEM)	139
9.3. Classification Accuracy.....	141
9.4. Reliability for Fixed Forms (Science)	142
Section 10: Validity	145
10.1. Intended Purposes and Uses of Test Scores	145
10.2. Sources of Validity Evidence	146
10.3. Evidentiary Validity Framework	147
10.4. Interpretive Argument Claims	150
10.5. NSCAS Validity Argument	151
References.....	153
Appendix A: Data Review Cheat Sheet	157
Appendix B: Summary of <i>P</i> Values by Item Type	162
Appendix C: Summary of Item-Total Correlations by Item Type.....	167
Appendix D: Achievement Level Distributions & Scale Score Descriptive Statistics by Demographics	173
Appendix E: Marginal Reliability by Demographics	178
Appendix F: Scatterplots for Scale Score CSEM.....	183
Appendix G: Alignment Study	188

List of Tables

Table 1.1. Schedule of Major Events for the Spring 2023 Administration.....	16
Table 2.1. NSCAS Growth in 2022–2023	20
Table 2.2. Number of Items and Points Per Test.....	22
Table 2.3. Online Item Types.....	25
Table 2.4. Data Review Flagging Criteria—Multiple-Choice Items	38
Table 2.5. Data Review Flagging Criteria—Non-Multiple-Choice Items	38
Table 2.6. Data Review Results.....	39
Table 3.1. User Roles and Responsibilities	41
Table 3.2. Accommodations and Universal Features	43
Table 3.3. Partner Support Communication Options	48
Table 3.4. Number of NSCAS Cases to Partner Support in 2022–2023	49
Table 4.1. Attemptedness Rules for Scoring	50
Table 4.2. MLE Scoring	50
Table 4.3. Score Range (LOSS and HOSS) for NSCAS Scale Score and Estimated RIT Score	51
Table 4.4. Achievement Level Descriptions for ELA, Mathematics and Science.....	51
Table 4.5. Reporting Categories	52
Table 4.6. Not-Tested Codes (NTCs)	53
Table 5.1. Blueprint Constraint Accuracy by Reporting Category—Fall Simulations.....	60
Table 5.2. Blueprint Constraint Accuracy by Reporting Category—Fall Engine Evaluation	61
Table 5.3. Blueprint Constraint Accuracy by Reporting Category—Winter Simulations	62
Table 5.4. Blueprint Constraint Accuracy by Reporting Category—Winter Engine Evaluation...	63
Table 5.5. Blueprint Constraint Accuracy by Reporting Category—Spring Simulations	64
Table 5.6. Blueprint Constraint Accuracy by Reporting Category—Spring Engine Evaluation...	65
Table 5.7. Item Exposure Rates—Fall Simulations	67
Table 5.8. Item Exposure Rates—Fall Engine Evaluation	67
Table 5.9. Item Exposure Rates—Winter Simulations	68
Table 5.10. Item Exposure Rates—Winter Engine Evaluation	68
Table 5.11. Item Exposure Rates—Spring Simulations.....	69
Table 5.12. Item Exposure Rates—Spring Engine Evaluation	70
Table 5.13. Mean Bias of the NSCAS Ability Estimation (True–Estimated)—Fall Simulations...	71
Table 5.14. Mean Bias of the NSCAS Ability Estimation (True–Estimated)—Winter Simulations	72
Table 5.15. Mean Bias of the NSCAS Ability Estimation (True–Estimated)—Spring Simulations	74
Table 5.16. Score Precision & Reliability, Items Contributing to NSCAS—Fall Simulations.....	76
Table 5.17. Score Precision & Reliability, Items Contributing to NSCAS—Fall Engine Evaluation	77
Table 5.18. Score Precision & Reliability, Items Contributing to NSCAS—Winter Simulations..	78
Table 5.19. Score Precision & Reliability, Items Contributing to NSCAS—Winter Engine Evaluation.....	79
Table 5.20. Score Precision & Reliability, Items Contributing to NSCAS—Spring Simulations..	80
Table 5.21. Score Precision & Reliability, Items Contributing to NSCAS—Spring Engine Evaluation.....	81

Table 5.22. SEM by Deciles for NSCAS Scores—Fall Simulations	83
Table 5.23. SEM by Deciles for NSCAS Scores—Fall Engine Evaluation	83
Table 5.24. SEM by Deciles for NSCAS Scores—Winter Simulations	84
Table 5.25. SEM by Deciles for NSCAS Scores—Winter Engine Evaluation	84
Table 5.26. SEM by Deciles for NSCAS Scores—Spring Simulations	84
Table 5.27. SEM by Deciles for NSCAS Scores—Spring Engine Evaluation	85
Table 5.28. General Population Demographic Distribution	87
Table 5.29. Demographic Distribution by Form—Grade 5 (Simulation)	87
Table 5.30. Demographic Distribution by Form—Grade 8 (Simulation)	88
Table 5.31. Demographic Distribution by Form—Grade 5 (Engine Evaluation)	90
Table 5.32. Demographic Distribution by Form—Grade 8 (Engine Evaluation)	91
Table 6.1. Number of Students Included in the Psychometric Analyses	93
Table 6.2. Summary of <i>P</i> Values—Operational Items	95
Table 6.3. Summary of <i>P</i> Values—Field-Test Items	96
Table 6.4. Summary of Item-Total Correlations—Operational Items	98
Table 6.5. Summary of Item-Total Correlations—Field-Test Items	99
Table 6.6. Flagging Criteria for MC Items	100
Table 6.7. Flagging Criteria for Non-MC Items	100
Table 6.8. Items to Be Suppressed	101
Table 6.9. Focal and Reference Groups for Gender- and Ethnicity-Based DIF	102
Table 6.10. LR DIF Categories	104
Table 6.11. MH DIF Categories for Dichotomous Items	105
Table 6.12. MH DIF Categories for Polytomous Items	105
Table 6.13. LR DIF Results—Field-Test Items (ELA/Mathematics)	106
Table 6.14. LR UIDIF Results—Field-Test Items (ELA/Mathematics)	108
Table 6.15. MH DIF Results—Field-Test Items (Science)	109
Table 6.16. Summary of IRT Item Statistics—Operational Items	110
Table 6.17. Summary of IRT Item Statistics—Field-Test Items	110
Table 6.18. Score Range (LOSS and HOSS) and Assigned Score	112
Table 6.19. Conversion of Theta to Scale Scores	113
Table 7.1. Final Approved Cut Scores	118
Table 8.1. Number of Students Tested by Demographics and Accommodations—Grade 3	119
Table 8.2. Number of Students Tested by Demographics and Accommodations—Grade 4	120
Table 8.3. Number of Students Tested by Demographics and Accommodations—Grade 5	121
Table 8.4. Number of Students Tested by Demographics and Accommodations—Grade 6	122
Table 8.5. Number of Students Tested by Demographics and Accommodations—Grade 7	123
Table 8.6. Number of Students Tested by Demographics and Accommodations—Grade 8	124
Table 8.7. Number of Students Tested by Administration Mode	125
Table 8.8. Testing Time in Minutes—ELA	127
Table 8.9. Testing Time in Minutes—Mathematics	128
Table 8.10. Testing Time in Minutes—Science	129
Table 8.11. Achievement Level Distributions	130
Table 8.12. Scale Score Descriptive Statistics	131
Table 8.13. Reporting Category Correlations	132
Table 8.14. Reporting Category Disattenuated Correlations	133
Table 8.15. Correlation and Descriptive Statistics of NSCAS and MAP Growth Scores	135

Table 9.1. Marginal Reliability of Scale Scores.....	136
Table 9.2. Marginal Reliability—Variance	138
Table 9.3. CSEMs at the Proficient Cut Scores	139
Table 9.4. Mean CSEMs by Decile	140
Table 9.5. Classification Accuracy by Achievement Level.....	141
Table 9.6. Cronbach’s Alpha (Internal Consistency) by Demographics for Science Fixed Forms	143
Table 10.1. Sources of Validity Evidence for Each NSCAS Test Purpose	147
Table 10.2. Sources of Validity Evidence Based on Test Content.....	147
Table 10.3. Sources of Validity Evidence Based on Response Processes	148
Table 10.4. Sources of Validity Evidence Based on Internal Structure.....	149
Table 10.5. Sources of Validity Evidence Based on Relations to Other Variables	150
Table 10.6. Interpretive Argument Claims—Evidence to Support Essential Validity Elements	151

List of Figures

Figure 1.1. Principled Test Design Process to Support Test Score Interpretations and Uses....	19
Figure 2.1. Test Development Process.....	20
Figure 2.2. Range ALD Example: ELA Grade 3	28
Figure 2.3. Item Development Plan	32
Figure 3.1. NSCAS Growth Platform Student Login Screen.....	40
Figure 5.1. Adaptive Engine Overview.....	56
Figure 5.2. Student-Specific Plan Approach	57

List of Abbreviations

Below is a list of abbreviations that appear in this technical report.

ALD	achievement level descriptor
CCC	Crosscutting Concept
CCR	College and Career Readiness
DCI	Disciplinary Core Idea
DIF	differential item functioning
DOK	Depth of Knowledge
DRC	Data Recognition Corporation
EDS	Educational Data Systems
ELA	English language arts
ELL	English language learner
ESEA	Elementary and Secondary Education Act
ESC	Education Strategy Consulting
ESU	educational service unit
ETS	Educational Testing Service
FT	field test
HL	horizontal linking
ID	Item-Descriptor
ISR	Individual Student Report
IEP	Individualized Education Plan
IRT	item response theory
IWW	item writer workshop
LOSS	lowest obtainable scale score
MC	multiple-choice
MLE	maximum likelihood estimation
NCCRS-S	Nebraska College and Career Ready Standards for Science
NCLB	No Child Left Behind
NDE	Nebraska Department of Education
NeSA	Nebraska State Accountability
NSCAS	Nebraska Student-Centered Assessment System
NTC	not-tested code
OIB	ordered item book
OP	operational
PP	paper-pencil
RAEL	Recently Arrived Limited English Proficient
SD	standard deviation
SEM	standard error of measurement
SEP	Science and Engineering Practice
SFTP	Secure File Transfer Protocol
STARS	School-based Teacher-led Assessment and Reporting System
TAC	Technical Advisory Committee
TCC	test characteristic curve
TEI	technology-enhanced item
TOS	Table of Specifications
TTS	text-to-speech
UAT	user acceptance testing

UDL Universal Design for Learning
VL vertical linking
VOIP Voice Over Internet Protocol

Executive Summary

This technical report documents the processes and procedures implemented to support the 2022–2023 Nebraska Student-Centered Assessment System (NSCAS) Growth in English language arts (ELA), mathematics, and science assessments by NWEA® under the supervision of the Nebraska Department of Education (NDE). The technical report shows how the processes, methods applied, and results relate to the issues of validity and reliability and to the *Standards for Educational and Psychological Testing* (AERA et al., 2014). Below is a high-level summary of each section in the technical report.

Section 1: Introduction

In Fall and Winter 2022–2023, the NSCAS assessments were administered in ELA and mathematics for grades 3–8. In Spring 2022–2023, the NSCAS assessments were administered in English language arts (ELA) and mathematics for grades 3–8 and in science for grades 5 and 8. The purposes of the NSCAS assessments are to measure and report Nebraska students' depth of achievement regarding the Nebraska College and Career Ready Standards; to determine if student achievement demonstrates sufficient academic proficiency to be on track for achieving college readiness; to measure students' annual progress toward college and career readiness; to inform teachers how student thinking differs along different areas of the scale, as represented by the range achievement level descriptors (RALDs), as information to support instructional planning; and to assess students' construct-relevant achievement in ELA, mathematics, and science for all students and subgroups of students.

Section 2: Test Design and Development

The Nebraska College and Career Ready Standards have been adopted by the Nebraska State Board of Education for ELA in 2021, mathematics in 2022, and science in 2017, respectively. The design of the NSCAS assessments is based on a principled approach to test design in which the evidence needed to draw a conclusion about where a student is in their learning of content is made explicit in the RALDs, and items are developed according to those pieces of evidence. To fully represent the constructs being assessed by NSCAS to determine if students are ready for college and careers, the adherence to specifications, the common interpretations of the standards, and an agreed-upon approach for cognitive complexity across all item types were closely monitored during item, passage, and test development.

Section 3: Test Administration and Security

The Spring 2023 NSCAS testing window was scheduled from April 3–May 12, 2023. The tests were administered online, with paper-pencil versions available as an accommodation. Appropriate accommodations and universal features were provided, and test security was adhered to throughout the entire test-administration process for both online and paper-pencil testing. User acceptance testing (UAT) was conducted prior to the operational administration to make sure the technology and item functionality were working properly.

Section 4: Scoring and Reporting

The adaptive online ELA and mathematics assessments were administered via NWEA's adaptive constraint-based engine (known as Cadabra). All tests were scored using maximum likelihood estimation (MLE) scoring. All steps of scoring went through a quality control process. Score reports were prepared at the individual student, school, district, and state levels.

Section 5: Adaptive Test Engine

During the assessment, NWEA's Cadabra engine administers items adaptively to match the ability level of each individual student. It has two stages of consideration as it selects the next item that conforms to the blueprint while providing the maximum information about the student based on the student's momentary ability estimate: the item selection for multiple feasible student-specific plans (SSPs), followed by choosing the complete SSP that maximizes guideline adherence and information. Pre-administration simulations and a post-administration evaluation study were conducted. Overall, NWEA's adaptive engine performed as expected.

Section 6: Psychometric Analyses

The following post-administration analyses were conducted for the ELA, mathematics, and science assessments: classical item analyses, including item difficulty (p value), item discrimination, and item suppression; differential item functioning (DIF) based on gender and ethnicity; and item response theory (IRT) calibration.

Section 7: Standard Setting

In July 2023, a standard setting meeting took place for ELA and mathematics, and a standards validation meeting took place for science. ACS Ventures was contracted to conduct the ELA and math standard setting and the science standards validation. ACS Ventures worked with panels of Nebraska educators through the process of recommending two cut scores to be used to distinguish the three achievement levels (i.e., *Developing*, *On Track*, *Advanced*). The purpose of the standard setting was to set new cut scores for mathematics and ELA, whereas the purpose of the cut score review (standards validation meeting) was to validate the existing cut scores for science.

Section 8: Test Results

More than 20,000 students were assessed in each grade and content area. Of those students across grades, half were males, half were females, two-thirds were white, and about one-fifth were Hispanic. Most students finished the tests within 120 minutes. The percentages of students at *Developing* are 37–46%, 34–42%, and 23–35% for ELA, mathematics, and science, respectively. Correlation coefficients between MAP Growth and NSCAS scores for students who took both tests in Spring 2023 were calculated. In general, these high correlations indicate that the relationship between MAP Growth and NSCAS test scores is strong, which can be considered validity evidence based on other variables.

Section 9: Reliability

The reliability/precision of the 2023 NSCAS assessments was examined through analysis of measurement error in simulated and operational conditions, including adaptive engine (Cadabra) score precision and reliability, marginal reliability, conditional standard error of measurement (CSEM), and Cronbach's alpha and standard error of measurement (SEM) for fixed forms. Marginal reliability estimates for the total scores are well above 0.80, which is typically considered the minimally acceptable level of reliability. The overall CSEM is consistent with reliability results. The classification accuracy results suggest that accurate classifications are being made for Nebraska students on the NSCAS assessments.

Section 10: Validity

Validating a test-score interpretation is not a quantifiable property but an ongoing process, beginning at initial conceptualization of the construct and continuing throughout the entire

assessment process. As this technical report progresses, it covers the different phases of the testing cycle, as well as the procedures and processes applied to the NSCAS assessments. This section revisits phases and summarizes relevant evidence and a rationale in support of any test-score interpretations and intended uses based on the *Standards for Educational and Psychological Testing* (AERA et al., 2014). The validity argument begins with a statement of the assessment's intended purposes followed by the evidentiary framework, where available validity evidence is provided to support the argument that the test actually measures what it purports to measure (SBAC, 2016).

While NSCAS assessments offer the additional benefit of reporting category scores that indicate directions for gaining further instructional information through the interim system or classroom observation, scores based on NSCAS are as equally reliable and valid as the traditional end-of-year assessment due to the following factors: First, NSCAS assessments go through the same rigorous psychometric analyses, such as test reliability, classification accuracy, CSEMs, test information, DIF, and a convergent validity check, and the results we have so far strongly support the reliability and validity claims of NSCAS assessments. In addition, the test-development process ensures validity of the intended test-score interpretations provided through the Reporting ALDs and scale scores. Last but not least, per the *Standards* (AERA et al., 2014), NSCAS assessments are aligned to grade-level content, and their test scores are suitable for use in accountability systems as a result of a robust development process of table of specifications (TOS), passage and item specifications, and achievement level descriptors (ALDs).

Section 1: Introduction

The purpose of this technical report is to summarize the design, development, administration, technical processes, and results of the Nebraska Student-Centered Assessment System (NSCAS) Growth assessments to support test users in evaluating the intended purposes, uses, and interpretations of the test scores. For 2022–2023, the through-year model was used in English language arts (ELA) and mathematics for grades 3–8, which were administered for Fall, winter, and spring; spring assessments include science for grades 5 and 8. NSCAS was designed by the state of Nebraska with support from its vendor, NWEA, to meet the requirements of the *Standards for Educational and Psychological Testing* (AERA et al., 2014) and federal peer review requirements (USDE, 2018) with an emphasis on using a principled assessment-design process.

1.1. NSCAS Overview

NSCAS is a statewide assessment system that embodies Nebraska's holistic view of students and helps them prepare for success in postsecondary education, career, and civic life. It uses multiple measures throughout the year to provide educators and decision-makers at all levels with the insights they need to support student learning. The NSCAS assessment, developed specifically for Nebraska and aligned to the state content area standards, is the assessment system's criterion-referenced measure designed for the Nebraska student population in grades 3–8.

The NSCAS assessments were administered online. They included a variety of item types, including multiple-choice and technology-enhanced items. Student scores were reported as composite scale scores and achievement levels. The ELA and mathematics assessments were administered online using an adaptive design, whereas science was administered as fixed forms. Students taking the NSCAS assessments were placed into one of the following achievement levels based on their final test scores:

- *Developing*
- *On Track*
- *Advanced*

Items for the ELA and mathematics tests were aligned to the 2014 and 2015 College and Career Ready Standards, respectively, and came from the item bank that the Nebraska Department of Education (NDE) and Nebraska educators have built over the years, including items field tested in Spring 2018 through Spring 2022. The spring tests also included previously and newly developed field-test items that will be added to the operational pool for the future, depending on the field-test data and data review. Content development for the new three-dimensional science assessment began in Summer 2018, with the pilot occurring in March 2019. A full-scale field test was also administered in Spring 2021 to gain feedback from Nebraska students on newly developed performance tasks. The new science assessments that were aligned to the Nebraska College and Career Ready Standards for Science (NCCRS-S; NDE, 2017) were administered in Spring 2022.

1.2. Background

From 2001 to 2009, Nebraska administered a blend of local and state-generated assessments called the School-based Teacher-led Assessment and Reporting System (STARS) to meet No Child Left Behind (NCLB) requirements. STARS was a decentralized local assessment system

that measured academic content standards in reading, mathematics, and science. The state reviewed every local assessment system for compliance and technical quality. NDE provided guidance and support for Nebraska educators by training them to develop and use classroom-based assessments. For accreditation, districts were also required to administer national norm-referenced tests. As a component of STARS, NDE administered one writing assessment annually in grades 4, 8, and 11. NDE also provided an alternate assessment for students severely challenged by cognitive disabilities.

Nebraska Revised Statute 79-760.03,¹ passed by the 2008 Nebraska Legislature, requires a statewide assessment of the Nebraska academic content standards for reading, mathematics, science, and writing in Nebraska's K–12 public schools. The new assessment system was named the Nebraska State Accountability (NeSA). NeSA replaced previous school-based assessments for purposes of local, state, and federal accountability and was phased in beginning with the 2009–2010 school year.

Through the 2015–2016 academic year, assessments in reading and mathematics were administered in grades 3–8 and 11; science was administered in grades 5, 8, and 11; and writing was administered in grades 4, 8, and 11. The 2015–2016 year was the final administration of the NeSA reading, mathematics, and science tests in grade 11. Nebraska adopted the ACT for high school testing in 2016–2017. NeSA-ELA tests were also implemented in Spring 2017, replacing NeSA reading.

NSCAS replaced the NeSA assessments beginning in 2017–2018. Spring 2022 was the fourth administration of the NSCAS ELA and mathematics assessments that were administered adaptively, whereas science continued to be administered as a fixed-form assessment. The new NSCAS science assessment, aligned to the NCCRS-S, was piloted in March 2019, with a full-scale field test administered in Spring 2021. Due to the COVID-19 pandemic, the Spring 2020 NSCAS administration was cancelled, delaying the timeline from an operational launch in Spring 2021 to Spring 2022.

To ensure a successful transition to a through-year assessment that capitalizes on the benefits of MAP Growth while also meeting the state requirements for identifying proficiency, a link was established between the NSCAS and MAP Growth scales.

1.3. Schedule of Major Events

Table 1.1 presents the major events that occurred for the 2023 NSCAS assessments, including the new science assessment. NDE involves educators throughout the development process to produce customized items and provide an invaluable professional-development opportunity, including item/task writing and review meetings and achievement level descriptor (ALD) reviews.

¹ <https://nebraskalegislature.gov/laws/statutes.php?statute=79-760.03>

Table 1.1. Schedule of Major Events for the Spring 2023 Administration

Event	Date(s)
Technical Advisory Committee Meeting	January 12, 2023
Mathematics Range ALD Workshop	February 27–March 3, 2023
Test Administration Training	July 26, July 28, August 9, and August 11, 2022, and February 22, 2023
Operational Test Window	April 3–May 5, 2023
Make-Up Test Window	May 8–May 12, 2023
District Review of Preliminary Data and Submission of Updates	May 15–May 19, 2023
ELA Alignment Study Workshop	July 24–July 28, 2023
ELA Standard Setting	July 25–27, 2023
Mathematics Standard Setting	July 25–27, 2023
Science Standards Validation	July 27, 2023
Delivery of Individual Student Reports (ISRs)	September 18, 2023
Math Data Review	October 18, 2023
Science Data Review	October 18, 2023
ELA Data Review	October 26, 2023

1.4. Building a Validity Argument

NSCAS assessments have been developed based on a principled approach to test design that centers around range achievement level descriptors (RALDs) and conceptualizing test-score use as part of a broader solution to achieve important outcomes for test users. The evidence needed to draw a conclusion about where a student is in their learning of content is made explicit in the RALDs, and items are developed according to those evidence pieces (Huff et al., 2016; Egan et al., 2012; Schneider & Johnson, 2019). This approach builds validity evidence into the design from the very beginning of the process, which is especially important when the assessments are intended to support interpretations regarding how student learning grows more sophisticated over time (Pellegrino et al., 2016). The purposes of a test design centered in RALDs include the following:

- To show how students increase in their reasoning with specific content across achievement levels to support collecting purposeful evidence of what mastery of college and career readiness means
- To support teachers in making more accurate inferences about what students know and can do

RALDs demonstrate how skills become more sophisticated as achievement and performance increase (Schneider et al., 2013). Such skill advancement is often related to increases in content difficulty and reasoning complexity and a reduction in the supports required for students to demonstrate what they know within a task or item. This use of RALDs helps teachers interpret the student work evidence to better identify where a student is in their learning and what they need next. Using a principled test-design process supports teachers in better understanding that a single standard has easier and more-difficult representations and that the goal of instruction is to support the development of cognitive skills in addition to content-based skills.

NDE took a balanced approach to the development process of the NSCAS assessments. Beginning with Policy ALDs, which are high-level expectations of student achievement within each achievement level across grades, NWEA (with input from Nebraska educators) developed Range ALDs, which define within-standard learning progressions that describe the knowledge and skills students at each achievement level can likely demonstrate. They describe the current stage of learning within the standard and explicate observable evidence of achievement, demonstrating how skills change and become more sophisticated across achievement levels for each standard.

Range ALD progressions were added to the item specifications in the item pool and used to support field test item development. After the test blueprint was finalized, the updated item pool was used to run simulations of the computer adaptive test (CAT) engine (Cadabra) in preparation for the student test event or fixed-form assessments.

Following test administration, cut scores for the achievement levels are defined during a Cut Score Workshop, or standard setting. Using evidence from the test scale and the adopted final cut scores, finalized versions of the Range ALDs were created and linked to the Reporting and Policy ALDs. Content interpretations were finalized after the standard setting and are used to support item specifications to ensure a stable, comparable construct over time.

With a principled approach to test design, RALDs may be viewed as the score interpretation, or the construct-interpretive argument described by Kane (2013). For RALDs to be the foundation of test-score interpretation, they should reflect more complex knowledge, skills, and abilities (KSAs) as the achievement levels increase (Schneider et al., 2013). As such, NDE developed RALDs to articulate the following:

- The observable evidence teachers and item developers should elicit to draw conclusions about a student's current level of performance
- What that evidence looks like when students are in different stages of development, represented by different achievement levels
- How the student is expected to grow in reasoning and content-skill acquisition across achievement levels within and across grades

Using RALDs, the NSCAS item bank has been aligned to the standards, represents the intended blueprint, and provides supports for students at all levels of proficiency within on-grade content. RALDs were developed in an iterative manner based on feedback from educators (Plake et al., 2010), with the final RALDs providing the interpretive argument regarding what test scores mean. By developing RALDs this way, Nebraska is communicating how standards are interpreted for assessment purposes, how tasks can align to a standard but not be of sufficient difficulty and depth to represent mastery, and what growth on the test-score continuum represents.

1.5. Intended Purposes and Uses of Test Results

Building a validity argument begins with identifying the purposes of the assessment and the intended uses of its test scores. The following are purposes of the NSCAS assessments:

1. To measure and report Nebraska students' depth of achievement regarding the Nebraska College and Career Ready Standards

2. To determine if student achievement demonstrates sufficient academic proficiency to be on track for achieving college readiness
3. To measure students' annual progress toward college and career readiness
4. To inform teachers how student thinking differs along different areas of the scale, as represented by the Range ALDs, as information to support instructional planning
5. To assess students' construct-relevant achievement in ELA, mathematics, and science for all students and subgroups of students

Ultimately, how test scores are used is determined by Nebraska educators. However, some intended uses of the NSCAS test results include the following:

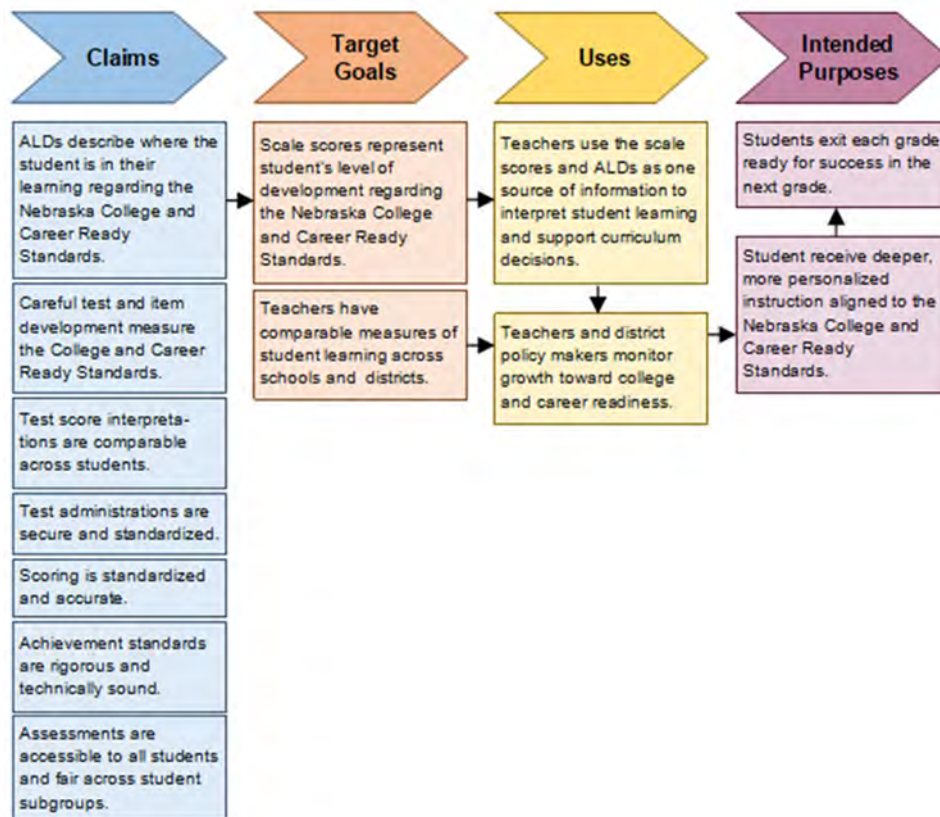
- To supplement teachers' observations and classroom assessment data
- To improve the decisions teachers make about sequencing instructional goals, designing instructional materials, and selecting instructional approaches for groups and individuals
- To identify individuals for summer school and other remediation programs
- To gauge and improve the quality of education at the class, school, system, and state levels throughout Nebraska
- To assess the performance of a teacher, school, or system in conjunction with other sources of information

1.6. Theory of Action

A theory of action is a tool that connects test users and their needs to decisions made during test design and development. In other words, it connects the design of the assessment (such as decisions about what evidence to collect and how to provide that evidence) to claims that test-score interpretation and use contribute to a positive solution to the broader problem for the test user. Figure 1.1 presents the theory of action for the NSCAS system. The ultimate intended purpose of NSCAS is to have students exiting each grade ready for success in the next grade. Evidence to determine if the assessment system is supporting its intended purposes across time may include the following:

1. Does Nebraska have increases in percentages of students who are *On Track* for college and career readiness?
2. Are students who are at or above *On Track* in one year likely to be *On Track* or above the following year?
3. Are students who are at or above *On Track* across time likely to be identified as *On Track* on an assessment of college or career readiness when scores are matched?

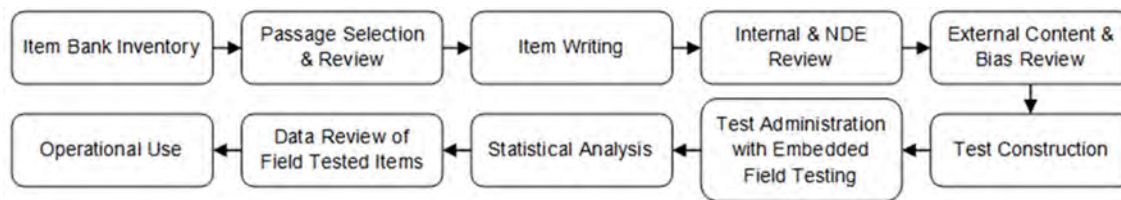
Figure 1.1. Principled Test Design Process to Support Test Score Interpretations and Uses



Section 2: Test Design and Development

This section describes the test design and development processes for the 2022–2023 NSCAS assessments. As Nebraska transitioned to an adaptive administration for ELA and mathematics in 2017–2018, the need to build a large, robust item bank was a key requirement, as was the development of new scales concurrent with the development of RALDs. Development of an item bank to sufficiently support the science assessments continued throughout 2022 in order to have enough content available to populate field-test slots in the Spring 2023 assessments. Items were written by educators in an item writing workshop (IWW). Once initial item development was completed, all items were taken to content and bias review meetings with Nebraska educators. Items that survived these meetings were considered for the field-test pool. Content development for the new three-dimensional science assessment began in Summer 2018, with the pilot occurring in March 2019, followed by the full-scale field test in Spring 2021. Figure 2.1 outlines the general steps taken to develop the passages and items.

Figure 2.1. Test Development Process



2.1. Test Designs

Table 2.1 summarizes the versions of the NSCAS Growth assessments available for 2023. Table 2.2 presents the number of items and points possible.

Beginning in 2022–2023, the fall and winter mathematics assessments were redesigned for more adaptivity (to be more similar to MAP Growth in that regard), and the summative blueprint is no longer strictly enforced. Therefore, additional flexibility for mathematics does not guarantee that all students will satisfy the 27-item summative blueprint.

The operational test was slightly longer for spring, having a total of 45 items, while the winter test had a total of 40 items.

Table 2.1. NSCAS Growth in 2022–2023

Content Area & Grade(s)	Available Assessments ^a				
	Online	PP	Spanish Online	Spanish PP	Breach
Fall/Winter					
ELA 3–8	Adaptive (40 items total per grade, 40 OP/DO items)	One form per grade (40 OP items)	Fixed (translation of PP form)	Same form as Spanish online	N/A
Mathematics 3–8	Adaptive (40 items total per grade, 40 OP/DO items)	One form per grade (40 OP items)	Fixed (translation of PP form)	Same form as Spanish online	N/A

Content Area & Grade(s)	Available Assessments ^a				
	Online	PP	Spanish Online	Spanish PP	Breach
Spring					
ELA 3–8	Adaptive (45 items total per grade, 38 OP/DO items and 7 FT items)	One form per grade (40 OP items)	Fixed (translation of PP form)	Same form as Spanish online	Winter PP form
Mathematics 3–8	Adaptive (45 items total per grade, 44 OP/DO items and 1 FT item)	One form per grade (40 OP items)	Fixed (translation of PP form)	Same form as Spanish online	Winter PP form
Science 5	20 forms (31 OP items and 6–10 FT items per form)	One form per grade (31 OP items and 2 FT items)	Fixed (translation of PP form)	Same form as Spanish online	NA
Science 8	20 forms (30 OP items and 8–11 FT items per form)	One form per grade (30 OP items and 3 FT items)	Fixed (translation of PP form)	Same form as Spanish online	NA

^a OP = operational; DO = diagnostic operational; PP = paper-pencil; FT = field test.

Table 2.2. Number of Items and Points Per Test

Grade	Adaptive						Fixed					
	Total	NSCAS Scores		RIT Scores		FT	Total	NSCAS Scores		RIT Scores		FT
	Items	Items	Points	Items	Points	Items	Items	Items	Points	Items	Points	Items
ELA (Fall)												
3	40	27–30	30–33	33–34	35–41	0	40	40	45	33	36	0
4	40	27–30	30–33	33–34	35–41	0	40	40	45	33	34	0
5	40	27–30	29–33	33–34	35–41	0	40	40	46	33	36	0
6	40	27–30	29–33	33–32	35–41	0	40	40	45	33	34	0
7	40	27–30	29–33	33–34	35–41	0	40	40	45	33	34	0
8	40	27–30	30–33	33–34	35–41	0	40	40	45	33	34	0
Mathematics (Fall) ^a												
3	40	27	31–35	44	44–48	0	40	40	43	40	43	0
4	40	27	31–35	44	44–48	0	40	40	44	40	44	0
5	40	27	31–35	44	44–48	0	40	40	44	40	44	0
6	40	27	31–35	44	44–48	0	40	40	44	40	44	0
7	40	27	31–35	44	44–48	0	40	40	43	40	43	0
8	40	27	31–35	44	44–48	0	40	40	44	40	44	0
ELA (Winter)												
3	40	27–30	30–38	33–34	36–40	0	40	40	46	29	33	0
4	40	27–30	31–38	33–34	35–40	0	40	40	46	30	34	0
5	40	27–30	32–38	33–34	36–40	0	40	40	47	30	35	0
6	40	27–30	32–38	33–32	36–40	0	40	40	45	30	33	0
7	40	27–30	31–36	33–34	36–40	0	40	40	44	31	34	0
8	40	27–30	30–36	33–34	36–41	0	40	40	45	30	33	0
Mathematics (Winter) ^a												
3	40	27	31–35	44	44–48	0	40	40	44	40	44	0
4	40	27	31–35	44	44–48	0	40	40	44	40	44	0
5	40	27	31–35	44	44–48	0	40	40	44	40	44	0
6	40	27	31–35	44	44–48	0	40	40	44	40	44	0
7	40	27	31–35	44	44–48	0	40	40	44	40	44	0
8	40	27	31–35	44	44–48	0	40	40	44	40	44	0
ELA (Spring)												
3	45	27–30	27–38	31–32	31–42	7	40	40	47	29	34	0
4	45	27–30	27–38	31–32	31–42	7	40	40	51	31	40	0
5	45	27–30	27–38	31–32	31–42	7	40	40	45	29	33	0

Grade	Adaptive						Fixed					
	Total	NSCAS Scores		RIT Scores		FT	Total	NSCAS Scores		RIT Scores		FT
	Items	Items	Points	Items	Points	Items	Items	Items	Points	Items	Points	Items
6	45	27–30	27–38	31–32	31–42	7	40	40	46	30	35	0
7	45	27–30	27–38	31–32	31–42	7	40	40	48	29	34	0
8	45	27–30	27–38	31–32	31–42	7	40	40	46	30	34	0
Mathematics (Spring)												
3	45	27	31–35	44	48–52	7	40	40	46	40	46	0
4	45	27	31–35	44	48–52	7	40	40	48	40	48	0
5	45	27	31–35	44	48–52	7	40	40	46	40	46	0
6	45	27	31–35	44	48–52	7	40	40	47	40	47	0
7	45	27	31–35	44	48–52	7	40	40	46	40	46	0
8	45	27	31–35	44	48–52	7	40	40	46	40	46	0
Science (Spring)												
5	37–41	31	33	N/A	N/A	6–10	31	31	33	N/A	N/A	2
8	38–41	30	33	N/A	N/A	8–11	30	30	33	N/A	N/A	3

Note. FT = field test.

^a NDE requested that the fall and winter test models in mathematics be redesigned for more adaptivity (to be more similar to MAP Growth in that regard), and in the case of mathematics, the summative blueprint is no longer strictly enforced.

2.2. Academic Content Standards

As stated in Nebraska Revised Statute 79-760.01² that was effective as of August 30, 2015:³

“The State Board of Education shall adopt measurable academic content standards for at least the grade levels required for statewide assessment pursuant to section 79-760.03. The standards shall cover the subject areas of reading, writing, mathematics, science, and social studies. The standards adopted shall be sufficiently clear and measurable to be used for testing student performance with respect to mastery of the content described in the state standards. The State Board of Education shall develop a plan to review and update standards for each subject area every seven years. The state board plan shall include a review of commonly accepted standards adopted by school districts.”

On September 5, 2014, the Nebraska State Board of Education adopted Nebraska’s College and Career Ready Standards for ELA. On September 4, 2015, the Nebraska State Board of Education adopted Nebraska’s College and Career Ready Standards for Mathematics. On September 8, 2017, the Nebraska State Board of Education approved the NCCRS-S that were implemented in the Spring 2019 pilot administration and later in the full-scale field test in Spring 2021.

2.3. Blueprints

The 2023 NSCAS blueprints for ELA and mathematics are embedded in the Table of Specifications (TOS) that indicate the range of test items included for each standards indicator. The adaptive test is constrained to make sure each student receives items within the identified ranges. The 2023 adaptive forms were not an exact match to the TOS given the attributes of available items in the item bank. Future forms will adhere more closely to the TOS as more items become available. The ELA TOS for each grade is available online at <https://www.education.ne.gov/assessment/nscas-general-summative-assessment/nscas-english-language-arts-ela/>. The mathematics TOS for each grade is available online at <https://www.education.ne.gov/assessment/nscas-general-summative-assessment/nscas-mathematics/>. The blueprint for the new science assessment is available online at <https://www.education.ne.gov/wp-content/uploads/2022/08/NE-Science-Public-Blueprint-Final.pdf>. This document provides an expectation of the frequency of the DCIs, SEPs, and CCCs from the NCCRS-S. Each element from the DCIs, SEPs, and CCCs is assigned a frequency (i.e., frequent, infrequent, rare) that indicates how often the element will be assessed.

2.4. Item Types

Table 2.3 presents the item types available for the online ELA and mathematics adaptive tests. Tasks field tested in science include phenomena and a set of items (i.e., prompts) using that phenomena that may include all of the available item types.

² <https://nebraskalegislature.gov/laws/statutes.php?statute=79-760.01>

³ <https://www.education.ne.gov/contentareastandards/>

Table 2.3. Online Item Types

Item Type	Description
Multiple-Choice (Choice)	Students select one response from multiple options.
Multi-select (Choice Multiple)	Students select two or more responses from multiple options. Some multi-select items are also two-point items for which students can earn partial credit.
Hot Text	Students select a response from within a piece of text or a table of information (e.g., word, section of a passage, number, symbol, or equation), which highlights the selected text. Some hot text items are also two-point items for which students can earn partial credit.
Text Entry	Students input answers using a keyboard.
Composite	Students interact with multiple interaction types included within a single item. Students may receive partial credit for composite items.
Drag & Drop	Students select an option or options in an area called the toolbar and move or “drag” these options (e.g., words, phrases, symbols, numbers, or graphic elements) to designated containers on the screen. Drag-and-drop items can include a click-and-click functionality in which students select the option and select the container it goes into instead of physically dragging it.
Gap Match	A type of drag-and-drop item in which students select one or more answer options from the item toolbox and populate a defined area, or “gap.”
Graphic Gap Match	A type of drag-and-drop item in which students move one or more answer options from the toolbox and populate a defined area, or “gap,” that has been embedded within an image in the item response area.

2.5. Depth of Knowledge (DOK)

With a principled approach to test design based on RALDs, increases in cognitive processing complexity (e.g., DOK, difficulty, context) are intended to be embedded into evidence statements across achievement levels in a cogent way and to interact with content. In this way, the features of cognitive processing, content difficulty, and context interact to affect item difficulty. A principled approach to test design is intended to support the validity of inferences about the student’s stage of learning and the content validity of the assessment as a measure of student achievement. Under such a score-interpretation model, construction of test blueprints should eventually cease treating DOK as a separate blueprint constraint. Instead, DOK should be present as evidence embedded in a descriptor for an achievement level that supports interpretations regarding the stage of thinking sophistication the student is at during the time of the test event (in addition to other factors that may affect difficulty, such as supports in the item). The items found within each achievement level should match the ALDs. The degree of alignment of items to the assessment, a component of the evidence gathered to support a validity framework, should focus on the degree of concurrence in the DOK and content alignment of items within an achievement level to the associated RALDs.

To ensure that the NSCAS assessments include a deep pool of items that span a full range of cognitive levels and skills, each item in ELA and mathematics was evaluated and tagged with one of the following DOK levels (Webb, 1997). DOK Level 4: Extended Thinking items are not included because the tests do not contain any extended-response items or performance tasks.

- DOK 1: Recall
- DOK 2: Skill & Concepts
- DOK 3: Strategic Thinking

Items at DOK 2 and 3 require conceptual and/or inferential thinking. DOK 3 items typically demand that students analyze and synthesize concepts from various parts of a text or from the text as a whole. ELA passages demonstrate varying degrees of complexity to support students at all levels of achievement. Because the NSCAS ELA and mathematics tests are adaptive, the overall distribution of DOK for any given test event varies based on individual student achievement and other factors. In February 2018, the state adopted the policy that *Developing* items could be at or below the cognitive level of the standards, *On Track* items could be at the cognitive level of the standards, and *Advanced* (formerly *CCR Benchmark*) items could be at or above the cognitive level of the standards. This policy decision influenced the development of the RALDs and the review of field-test items.

2.6. ALD Development

The NSCAS ALDs were developed based on the following ALD development stages proposed by Egan et al. (2012) to correspond with the closely linked uses of ALDs in test development and score reporting. ALD development using this model is consistent with a construct-centered approach to assessment design (Messick, 1994).

1. Policy ALDs: High-level expectations of student achievement within each achievement level across grades, often defined by the state
2. Range ALDs: Detailed descriptions of each achievement level by grade that show students' increasing ability to apply practices and concepts
3. Reporting ALDs: Reflect student performance based on the final approved cut scores

2.6.1. Policy ALDs

The following Policy ALDs were developed to communicate the vision of what a test score is intended to represent, or where a student is in their learning regarding the content standards. When carefully crafted, Policy ALDs can be viewed as the assessment claim because they set the tone for how the content and cognitive demand are intended to be articulated along the test scale. The Nebraska Policy ALDs guide the establishment of the intended policy outcomes NDE desires for Nebraska students.

- *Developing* learners do not yet demonstrate proficiency in the knowledge and skills necessary at this grade level, as specified in the assessed Nebraska College and Career Ready Standards.
- *On Track* learners demonstrate proficiency in the knowledge and skills necessary at this grade level, as specified in the assessed Nebraska College and Career Ready Standards.
- *Advanced* learners demonstrate advanced proficiency in the knowledge and skills necessary at this grade level, as specified in the assessed Nebraska College and Career Ready Standards.

2.6.2. Range ALDs

Range ALDs provide the intended content-based interpretations of what test scores within an achievement level represent and explicate observable evidence of achievement, demonstrating how the skill changes and becomes more sophisticated across achievement levels for each

standard and achievement level on an assessment. Teachers can use Range ALDs to determine how students with different scores within different achievement levels may differ in their abilities. Range ALDs for ELA were developed in 2017 and reviewed by NWEA in 2018. Range ALDs for mathematics were developed in 2018, including an educator review in Spring 2018. Both ELA and mathematics Range ALDs were refined during the July 2018 standard setting and cut score review meetings. Range ALDs have also been generated for the new science assessment aligned to the NCCRS-S, beginning with an ALD workshop in May 2019.

2.6.2.1. ELA and Mathematics

To develop the ELA Range ALDs, educators at the July 2018 cut score review meeting used the ALDs from the original standard setting to develop a first draft. After the cut score review, NWEA reviewed the draft ALDs again, editing for consistency of language and clarity in a second draft and considering the final approved cut scores. Next, NWEA worked across grades to ensure a logical vertical progression and consistent language between the grades. Once a coherent and cohesive third draft was created, it was sent to NDE for review. NWEA implemented NDE's feedback and sent the resulting fourth draft back to NDE for an additional review and approval.

In 2022, NWEA worked with NDE to update the ELA Range ALDs to the newly adopted 2021 ELA standards. NWEA first provided NDE with a draft version of the ELA Range ALDs aligned to the new ELA standards. NDE reviewed and provided feedback, which NWEA implemented. Then, Nebraska ELA educators provided feedback during a five-day, virtual Range ALD workshop held June 6–10, 2022. NWEA implemented the educators' feedback and provided a final version to NDE for their review and approval. NDE signed off on this document, which is available online at <https://www.education.ne.gov/assessment/nscas-general-summative-assessment/nscas-english-language-arts-ela/>. This version of the ELA ALDs was used for the Spring 2023 assessment.

To develop the mathematics Range ALDs, an educator committee was convened in April 2018 to review a first draft. NWEA and NDE then engaged in an extensive revision process that involved several iterations of rework. The draft ALDs were brought to the July 2018 standard setting meeting, where they were reviewed and refined by educators based on the cut scores. After receiving the final approved cut scores, NWEA reconciled the ALDs based on item content, participant recommendations, and the final cut scores, consistent with recommended practice (Egan et al., 2012). Those edits were used to inform changes throughout the ALDs. These updates were shared with NDE for feedback. After receiving NDE's feedback, NWEA made the requested edits or responded to the posted questions. The files were then formatted and submitted to NDE. The final mathematics ALDs are available online at <https://www.education.ne.gov/assessment/nscas-general-summative-assessment/nscas-mathematics/>.

Figure 2.2 presents example Range ALDs for ELA grade 3 for the 2021 standards that were assessed in Spring 2023. The progression descriptor (i.e., *Developing, On Track, Advanced*) describes where a student is in their learning regarding the standard. Within a single expectation (e.g., LA.3.RP.1) can be ranges of content- and thinking-skill difficulty that describe different stages of reasoning.

Figure 2.2. Range ALD Example: ELA Grade 3

Indicator No.	Indicator Text	Developing	On Track	Advanced
		With a range of texts with text complexity commonly found in Grade 3, a student performing in Developing can likely	With a range of texts with text complexity commonly found in Grade 3, a student performing in On Track can likely	With a range of texts with text complexity commonly found at the intersection of Grade 3 and Grade 4, a student performing in Advanced can likely
Reading Prose and Poetry				
Central Ideas and Details				
Citing relevant and thorough textual evidence to support ideas, evaluate the development of themes or central ideas in grade-level literary texts.				
LA.3.RP.1	Identify the central message or lesson in a literary text and explain how key details support that idea.	Identify the central message or lesson in a literary text.	Identify the central message or lesson in a literary text and explain how key details support that idea.	Analyze the central message or lesson in a literary text and explain how key details support that idea.
LA.3.RP.2	Explain how characters respond to major events and challenges in a literary text.	Identify the major events and/or challenges that characters face in a literary text.	Explain how characters respond to major events and challenges in a literary text.	Analyze how characters respond to major events and challenges in a literary text, drawing on specific details such as a character's thoughts, words, or actions.
Author's Craft				
Citing relevant and thorough evidence to support ideas, evaluate the development and interaction of individuals, ideas, and events in grade-level literary text.				

Indicator No.	Indicator Text	Developing	On Track	Advanced
LA.3.RP.3	Determine and explain the point of view in a literary text.	Identify the narrator or speaker in a literary text.	Determine and explain the point of view in a literary text.	Analyze how the point of view influences a literary text.
LA.3.RP.4	Explain how sections of a literary text (e.g., chapters, scenes, stanzas) build on one another and contribute to meaning.	Identify and/or describe the sections of a literary text (e.g., chapters, scenes, stanzas).	Explain how sections of a literary text (e.g., chapters, scenes, stanzas) build on one another and contribute to meaning.	Analyze how sections of a literary text (e.g., chapters, scenes, stanzas) build on one another and evaluate which sections contribute most to meaning.

Source: [2021-Range-ALDs June-2022-2.xlsx \(live.com\)](#)

Nebraska's College and Career Ready Standards are organized so that each expectation level represents a specific skill or building block for problem solving. This could be a learning progression, but these indicators are in separate expectation levels. Therefore, how each indicator may be expected to increase in sophistication needs to be defined to support defining the test-score interpretations across achievement levels. Because the indicators are separate for these types of steps, the ALDs focus on other differentiating factors within each indicator to represent the progression of student knowledge and understanding of the specified skill. The ALDs also strive to preserve differentiation between the skills as they progress across grades. The following example shows where content limits (or conscious decisions about how content should increase in difficulty within an indicator) are used to differentiate items aligned with different achievement levels within an indicator, as well as across grades:

- Standard 3.N.1.b in grade 3 mathematics is about comparing whole numbers through the hundred thousands.
- The corresponding standard at grade 2 compares two three-digit numbers.
- The lower level of grade 3 continues the progression of the skill by comparing one three-digit number to a number between 1,000 and 100,000.
- The middle-level ALD then progresses to two numbers between 1,000, and 100,000.

The ALDs also differentiate between achievement levels through the presentation of information to the student or what supports are provided. In some cases, visual models are required at the lower level but not at the higher levels (provided the standard does not require visual models). The higher-level ALDs aim to require analysis of ELA and mathematics to better assess conceptual understanding and higher levels of cognitive processing while also staying true to the indicator. The definition of content across achievement levels in this way is critical to supporting the development of content aligned to the state indicators and expectations at the levels of specificity denoted by state's test blueprints in terms of numbers of items per indicator. All items under this framework align to the indicators, and the explicit manipulation of item features to support changes in item difficulty is consistent with the Range ALD development framework in which content difficulty, cognitive processing demands, and contextual features (such as scaffolding, visuals, and relationships with other standards) are explicitly built into the ALDs (Egan et al., 2012). While this approach is helpful in a fixed-form context, it is critical to item development for an adaptive assessment.

2.6.2.2. Science

Before task development began in Summer 2019 for the new science assessment, it was essential to first develop the ALDs that correspond to the *Developing*, *On Track*, and *Advanced* achievement levels to guide development. The science Range ALDs are intended to describe students' increasingly advanced three-dimensional reasoning on tasks that require students to apply and integrate SEPs and CCCs within and among the disciplines of science. The science ALDs are available online at https://www.education.ne.gov/wp-content/uploads/2022/08/NSCAS-Science-Summative-Achievement-Level-Descriptors-ALDs-Final_8.17.2022.pdf.

Nebraska's College and Career Ready Standards for Science (NCCRS-S) may be thought of as the broad content learning goals for students at each grade level that are intended to cue instruction in ways that emphasize active scientific reasoning, but there is complexity regarding how the standards are intended to be interpreted, taught, and assessed. Indicators found in the NCCRS-S are meant only to provide examples of ways the three-dimensional standards could be integrated on an assessment. Assessment tasks centered in the NCCRS-S are intended to

measure a novel indicator based on the intersection of the grade-level DCI, CCC, and SEP through a task-based claim (i.e., students are applying SEPs to make sense of task phenomena using the intended DCIs and CCCs). Because a task-based claim represents a novel indicator, indicators can and likely will vary across alternate test forms of a state assessment. The ALDs must do two things:

1. Be specific enough to describe increasingly advanced three-dimensional reasoning and the required evidence the assessment must have that is common across alternate tasks and alternate forms of the assessment
2. Be sufficiently generalized so that they may subsume novel indicators that change across time and, potentially, students

To accommodate these needs, NDE has determined that specific science content claims (i.e., DCIs) should not be the focus of the ALDs. Instead, the grade-level content articulated in the DCIs becomes the foundation for measuring complex integration of scientific reasoning (i.e., SEPs and CCCs) and setting up phenomena that can change across alternate test forms and, potentially, students. Therefore, Range ALDs must reflect the progression of proficiency claims regarding how SEPs and CCCs become more sophisticated as each achievement level increases. In particular, in a three-dimensional assessment that emphasizes active scientific reasoning, the on-grade content must be extended in some way to a different phenomenon or problem so that NDE can learn about student abilities in “reasoning like a scientist.”

The DCI dimension will be embedded into the phenomena-based tasks so that the ALDs represent the three dimensions, which is represented by a consistent header in the ALDs that addresses the phenomena. For each SEP, each achievement level will need to describe the evidence NDE expects to collect to infer that a student is within that achievement level. For example, the evidence for the *On Track* achievement level should articulate more advanced, explicit student behaviors compared with those articulated in the *Developing* achievement level.

Range ALDs define the expected differences in scientific reasoning, which is useful to teachers because it aligns the evidence to be collected for each achievement level with NDE’s vision for student performance in terms of mastery of the dimensions of the NCCRS-S. Dimensional progressions are described in *A Framework for K–12 Science Education* (National Research Council, 2012), a guiding document to the NCCRS-S and to the science ALD development process. Given that NDE expects to integrate these dimensions within tasks, the dimensions cannot be viewed as independent. One dimension can influence the complexity of another dimension and, therefore, the difficulty of prompts along the reporting scale. Thus, dimensions need to be integrated in the ALDs consistently in order to describe differences in student achievement. This also means that SEPs and CCCs need to be integrated consistently, even though the phenomena and problems used to measure those skills can vary.

2.6.3. Reporting ALDs

Reporting ALDs are provided at the overall score level and are optimally created after final cut scores are adopted following the standard setting procedure. Reporting ALDs represent the reconciliation of the Range ALDs with the final cut scores. The Range ALDs reflect a state’s initial expectation for student performance within an achievement level, whereas the Reporting ALDs reflect actual student performance based on the final approved cut scores. The Reporting ALDs define the appropriate inferences stakeholders may make based on the student’s test

score in relation to the final approved cut scores. Teachers are optimally given supportive information regarding how to interpret them to support formative practice.

2.7. Item Development

Item development for the 2022–2023 assessment administration was not required for math and ELA. Items field tested in 2022–2023 had already been developed in prior years. Science summative task and item development occurred during Summer 2022 in an item writing workshop.

2.7.1. Item Specifications

All items developed for the NSCAS assessments should align to one standard and should follow best practices for creating test items. The RALDs provide detailed information regarding each standard and how to assess student knowledge at different levels for each standard. Items should meet the level specified for each standard. Following best practices, including style, helps ensure that items are accurately measuring student knowledge at each level by focusing the items on construct-relevant information and presentation. The item specifications incorporate information from each source into a single file to provide a high-level overview for creating NSCAS test items.

There is a separate item-specifications document for each content area. Item specifications for both ELA and mathematics capture aspects such as those listed below and are reviewed at the start of each new development cycle to ensure accuracy. Item specifications for the new science assessment were based heavily on mathematics and are being updated collaboratively with NDE throughout the development process.

- General item writing guidelines in terms of overall content, item stems, item responses, style, and scoring rules
- Specific guidelines for using TEIs
- Specific standard information for grades 3–8
- Range ALDs

2.7.2. Science

An item-development plan was created based on an analysis of the Nebraska science task pool and how it could fulfill the grade 5 and grade 8 blueprints. Task alignments were selected to fill gaps across all the Next Generation Science Standards (NGSS) dimensions (SEPs, DCIs, and CCCs) as well as across the achievement level descriptors (ALDs). Combinations of dimensions were selected that would best help facilitate writing a compelling and coherent task set. This development plan takes into account teacher feedback and observed experiences and groups dimensions together that should lead to meaningful sense-making and exploration of a wide variety of phenomena. The item-development plan is outlined in Figure 2.3 below, as follows:

Figure 2.3. Item Development Plan

Grade	Focal NE Indicator	Focal DCI	Focal SEP	Focal ALD	Focal CCC
5	SC.5.3.1.A	PS1	ARG	ARG-5OT	EM
5	SC.5.3.1.B	PS1	MATH	MATH-5OT	SPQ
5	SC.5.3.1.B	PS1	MATH	MATH-5CCR	SPQ
5	SC.5.3.1.C	PS1	MATH	MATH-5OT	SPQ

Grade	Focal NE Indicator	Focal DCI	Focal SEP	Focal ALD	Focal CCC
5	SC.5.3.1.C	PS1	MAT	MATH-5CCR	SPQ
5	SC.5.3.1.D	PS1	INV	INV-5OT	CE
5	SC.5.3.1.D	PS1	INV	INV-5CCR	CE
5	SC.5.8.2.A	PS3	MOD	MOD-5OT	EM
5	SC.5.8.2.C	LS2	MOD	MOD-5OT	SYS
5	SC.5.8.2.C	LS2	MOD	MOD-5CCR	SYS
5	SC.5.11.3.A	PS2	ARG	ARG-5CCR	CE
5	SC.5.11.3.B	ESS1	ARG	ARG-5OT	SPQ
5	SC.5.13.4.A	ESS2	MOD	MOD-5OT	SYS
5	SC.5.13.4.B	ESS2	MATH	MATH-5OT	SPQ
5	SC.5.13.4.B	ESS2	MATH	MATH-5CCR	SPQ
5	SC.5.13.4.C	ESS3	INFO	INFO-5OT	SYS
5	SC.5.13.4.C	ESS3	AQ	AQ-5OT	SYS
5	SC.5.13.4.D	ETS1	DP	DP-5OT	SC
5	SC.5.13.4.D	ETS1	DP	DP-5CCR	SYS
5	SC.5.13.4.E	ETS2	DP	DP-5OT	EM
8	SC.8.1.1.A	PS2	CEDS	DS-8D	SYS
8	SC.8.1.1.B	ETS1	MOD	MOD-8CCR	SC
8	SC.8.1.1.D	PS2	AQDP	AQ-8OT	CE
8	SC.8.1.1.E	PS2	ARG	ARG-8OT	SYS
8	SC.8.1.1.F	PS2	INV	INV-8OT	CE
8	SC.8.1.1.F	PS2	INV	INV-8CCR	CE
8	SC.8.2.2.A	PS4	MATH	MATH-8OT	PAT
8	SC.8.2.2.C	PS4	INFO	INFO-8OT	SF
8	SC.8.4.3.A	PS3	DATA	DATA-8OT	SPQ
8	SC.8.4.3.B	PS3	MOD	MOD-8OT	SYS
8	SC.8.9.4.A	LS3	MOD	MOD-8OT	SF
8	SC.8.9.4.A	LS3	MOD	MOD-8CCR	SF
8	SC.8.9.4.B	LS4	INFO	INFO-8OT	CE
8	SC.8.10.5.A	LS4	DATA	DATA-8OT	PAT
8	SC.8.10.5.A	LS4	DATA	DATA-8CCR	PAT
8	SC.8.10.5.B	LS4	CEDS	CE-8OT	PAT
8	SC.8.10.5.C	LS4	CEDS	CE-8OT	CE
8	SC.8.11.6.B	ESS1	MOD	MOD-8D	SYS
8	SC.8.11.6.C	ESS1	DATA	DATA-8OT	SPQ
8	SC.8.14.7.A	ESS1	CEDS	CE-8OT	SPQ

Each science task contains the phenomena, text to support student thinking, any required graphics or tables, and the prompts to which the student must respond. The goal of each task is to evaluate student sense-making skills. During the workshop, the writers were guided in the vision of the NSCAS science assessment and began the development process by identifying a

phenomenon that met NDE's criteria (e.g., it is observable, accessible, engaging, and explainable using grade-level appropriate science core ideas). A phenomena or problem provides an overall context for the task. Writers then thought about the steps needed for students to make sense of the phenomenon and identified DCIs, SEPs and CCCs students would use in the sense-making process. A task was built by introducing the phenomenon in a scenario that was bimodal (e.g., it had text and graphics) in most cases followed by prompts that were minimally two-dimensional. When additional information was needed, it was presented with another mini-scenario. Each task had at least one three-dimensional prompt.

Nebraska teachers were recruited by NDE and brought together during Summer 2022 for a phenomena/item writer workshop. Teachers participated in the workshop to develop ten tasks for grade 5 and ten tasks for grade 8. Ten tasks per grade were also developed by NWEA subject-matter experts. The newly developed tasks and prompts were further refined during a review by a content and bias review committee, facilitated by NWEA, that consisted of NDE educators recruited by NDE who were not involved in writing the tasks for the grade they reviewed.

2.7.3. Item Retirement

Field-tested items are removed from the pool if they do not pass data review. Operational items are retired (i.e., removed) based on content and psychometric reviews of items flagged based on their item statistics and a set of flagging criteria after each administration. There is no limit to how many times an item can be used operationally. Items may also be re-field tested if deemed necessary (e.g., if an item required revisions for clarifications or if an item changed grades based on a new set of standards).

2.8. Content Alignment

To fully represent the constructs being assessed by NSCAS to determine if students are ready for college and careers, solid content alignment is critical. This was covered in several ways in prior developments for the items used in this administration, including adherence to specifications, common interpretations of the standards, and an agreed-upon approach for cognitive complexity across all item types.

2.8.1. Alignment and Adaptive Testing

Within an adaptive testing context, the documentation of content blueprint features and percentages of the items tagged to the blueprint features in the item pool become one evaluation tool used to frame alignment discussions. Both item pool structure and constraints used to establish the administration of items during test events support the definition of the construct for alignment purposes. Full test blueprints must be supportable for students in each achievement level. Therefore, an ideal item pool has similar percentages of items within each indicator by achievement level cell.

As RALDs were developed based on theories of how student thinking grows within the state's structure of state standards (and the evidence needed to support that conclusion), the characteristics of items depend on the student's stage of reasoning. As RALDs describe increases in student thinking and reasoning, test developers have a rationale regarding why a percentage of particular item types (e.g., technology-enhanced items) and DOK levels are necessary in the item bank, as well as the percentage of items that should be developed to particular levels of cognitive complexity within an item bank. Those decisions are driven based

on the construct-based evidence that should be collected and included in item specifications. These decisions are made within each indicator by achievement level cell.

Students who are in earlier stages of reasoning can be forced into more advanced cognitive levels with more difficult content when computer adaptive constraints force all students to receive a certain percentage of items at a particular DOK level. A fundamental development practice for the Range ALDs (Egan et al., 2012) is that DOK levels follow the indicator progression. While DOK may increase across achievement levels, the DOK level should not automatically increase with the achievement level. What may be required from a learning-theory perspective is that students have support accessing the standards, such as with visual supports demarcating a manipulation of an item context feature. They then may access the standards without the visual aids, followed by accessing the standards at a higher DOK level. Thus, if the item development is purposeful to the progression, DOK specifications are not required as a constraint, conditional that items are measuring what the RALDs say they are.

When item development is purposeful to a clearly defined construct, dictating a certain percentage of items at a particular DOK level will unintentionally route a student to items that provide less information about their current stage of thinking and reasoning with the content. Thus, from a student and item bank evaluation perspective, alignment processes must consider the specific item demands of the RALDs within an achievement level and ask independent judges if items align to a specific RALD within an achievement level. This can be done during external content reviews with educators. Subsequently, with the documented RALD matching of each item, the relationships among the achievement level categorizations, the item difficulty, and the degree of alignment can be used as evidence of alignment from a content validity perspective.

2.8.2. 2023 ELA Alignment Study

NWEA, on behalf of NDE, contracted with the Human Resources Research Organization (HumRRO) to evaluate the degree of alignment between the Nebraska Student-Centered Assessment System (NSCAS) in ELA and Nebraska's College and Career Ready Standards (NCCRS) in ELA. This virtual alignment study was held July 24–28, 2023, and gathered critical evidence to support inferences made about students' scores on the NSCAS in ELA.

Educators were recruited to serve on grade-level panels for grades 3–8. Panelists performed iterative steps for each item their panels reviewed. These steps included: 1) viewing secure test items, 2) entering independent ratings into a spreadsheet, 3) discussing independent ratings with other alignment workshop participants, and 4) determining final ratings for each item as a group. A full copy of the alignment study report can be found in Appendix G: Alignment Study.

As a result of the alignment study, NWEA has reviewed the feedback from HumRRO and will be taking some actions prior to and during the next round of development. These actions include:

- Completing an item bank analysis and identifying standards that do not have coverage or that have minimal coverage. These standards will be targeted during summer 2024 item development. (It is also worth noting that gaps identified during the summer 2023 alignment study will be filled by items that were developed in summer/fall 2023.)
- Developing more ELA items that align to DOK3 and ALD3.

- Discussing with NDE the possibility of revising the test specs to be at the standard level vs the sub-standard level. (This would solve the issue of there being more standards than items on the assessment.)

2.9. Universal Design

Ensuring that assessments are accessible to students with a variety of needs, including those with disabilities, is a critical part of item development. With a strong foundation in Universal Design for Learning (UDL), the assessments become engaging and accessible for all students. The NWEA content team ensures that each item is created with the principles of UDL in mind. These principles provide a framework for developing flexible items to support many kinds of learners and maximize options for assessments in order to provide multiple means of representation, action and expression, and engagement. Applying UDL principles to assessments helps reduce barriers and minimize irrelevant information from items so the assessment can show what each student knows.

2.10. Sensitivity and Fairness

NWEA takes seriously the task of creating items that are free from bias and sensitivity issues and are fair to all students. Items are revised to eliminate bias, sensitivity, and fairness issues—or rejected if an issue cannot be remedied through the revision process.

- **Bias:** This is defined as item content, unrelated to the concept or skill being assessed, that may unfairly influence a student's performance or an item construct that does not have equivalent meaning for all students
- **Sensitivity:** This can result if the experience of taking a test differs from the classroom experience in that students do not have the opportunity to discuss the material with a teacher or their peers. Sensitive content risks drawing students out of the testing experience by provoking negative emotional responses.
- **Fairness:** This is defined as the equitable treatment of all students during the assessment process. To make a test fair, test developers must work to eliminate any barriers that prevent students from understanding and interacting with item content in a manner that accurately demonstrates what they know or are able to do.

A successful item is free of bias and sensitivity issues and is accessible to all students. An item should NOT:

- Distract, upset, or confuse in any way
- Contain inappropriate or offensive topics
- Require construct-irrelevant knowledge or specialized knowledge
- Favor students from certain language communities
- Favor students from certain cultural backgrounds
- Favor students based on gender
- Favor students based on social economic issues
- Employ idiomatic or regional phrases and expressions
- Stereotype certain groups of people or behaviors
- Favor students from certain geographic regions
- Favor students who have no visual impairments
- Use height, weight, test scores, or homework scores as content or data in an item

There is not a hard and fast “list” of material that is potentially distracting or upsetting, but some topics are seldom appropriate for K–12 assessments, such as sexuality, illegal substances, illegal activities, excessive violence, discriminatory descriptions, death, grieving, catastrophes, animal neglect or abuse, and loss of a family member.

2.11. Test Construction (ELA and Mathematics)

The online adaptive tests were produced by selecting the item pools, building the test models that configured the engine and provided the constraints, running simulations, approving the results, and conducting user acceptance testing (UAT). The ELA and mathematics paper-pencil forms were created based on the blueprint and statistical guidelines.

2.11.1. Fixed-Forms

The ELA and mathematics fixed forms were created based on the blueprint and fixed-form construction specifications that included the following statistical guidelines:

- An absolute test characteristic curve (TCC) difference < 0.05
- A max of three items with a differential item functioning (DIF) flag of C- or C+
- A max of three items with an item-total correlation flag
- A max of three items with an omit rate $> 5\%$
- A max of three items with an item-total correlation for a distractor > 0.05
- A max of three items with a p value < 0.2 or > 0.9
- A max of three items with a p value for an answer key $<$ a distractor p value
- No items with an answer key item-total correlation $<$ the item-total correlation for a distractor
- No items with a negative item-total correlation

The content team also considered the following:

- Number of items per standard indicator
- Number of items at each level of cognitive complexity
- The balance between dichotomous and polytomous items
- The balance between multiple-choice and technology-enhanced items

Item selection was an iterative process between the psychometrics and content teams before being sent to NDE for review and approval.

2.11.2. MAP Growth Item Selection

For the through-year model, MAP Growth items were added to the item pool for diagnostic purposes. The NWEA content team reviewed the MAP Growth items and selected those that were aligned to NSCAS standards, conformed to NSCAS item specifications, and could contribute toward the test blueprint. Because a link was established between NSCAS ELA and MAP Growth Reading, only MAP Growth Reading items were considered; that is, MAP Growth Language Usage items were not included.

2.12. Data Review

Data review is the process of reviewing field-tested items for quality and appropriateness based on the results of statistical analysis of student responses. The review of content alignment and statistics of the Spring 2023 field-tested items occurred virtually in October/November 2023

between NDE and NWEA. Table 2.4. Data Review Flagging Criteria—Multiple-Choice Items and Table 2.5 present the data review flagging criteria for multiple-choice and non-multiple-choice items, respectively. Items were flagged based on these criteria and brought to the data review meeting for review and discussion.⁴ NWEA shared with participants the data review flagging criteria outlined in these tables.

Table 2.4. Data Review Flagging Criteria—Multiple-Choice Items

Statistic	Criterion	Indication
DIF of gender or ethnicity	C+ or C-	Potential bias toward a certain group of students
IRT difficulty or step parameters are extremely high	≥ 4.25	Probability of getting an item correct may require extremely high ability
Item fit statistics	< 0.7 or > 1.3	Poor fit
<i>P</i> value	< 0.20 or > 0.9	Very difficult item
<i>P</i> value for distractors	Distractor % $>$ key %	More students chose a distractor than the key
Item-total correlation	< 0.20	Poorly discriminating item
Item-total correlation for distractors	> 0.05	Poorly discriminating item
Omit rate	$> 5\%$	Unclear or very difficult item

Table 2.5. Data Review Flagging Criteria—Non-Multiple-Choice Items

Statistic	Criterion	Indication
DIF of gender or ethnicity	C+ or C-	Potential bias toward a certain group of students
IRT difficulty or step parameters are extremely high	≥ 4.25	Probability of getting an item correct may require extremely high ability
Item fit statistics	< 0.7 or > 1.3	Poor fit
Step parameters	Step 1 $>$ Step 2	Not a good separation of students into different stages of learning
Item-total correlation	< 0.1	Poorly discriminating item
Item-total correlation for score of 0	> 0.0	Poorly discriminating item
Item-total correlation for score of 1 $<$ item-total correlation for score of 0	–	Poorly discriminating item
Item-total correlation for score of 2	< 0.1	Poorly discriminating item
Item-total correlation for score of 2 $<$ item-total correlation for score of 1	–	Poorly discriminating item
Low student count for each score	$= 0$	No one got a certain score (e.g., no student got a score of 2)

Table 2.6 presents the data review results, including the number of field-test items included in the pool, the number of field-test items administered during the 2023 testing window, the number of field-test items included for data review, the number of rejected field-test items, and the number of accepted field-test items.

⁴ The summaries of item analyses are included in Section 6: of this technical report.

Table 2.6. Data Review Results

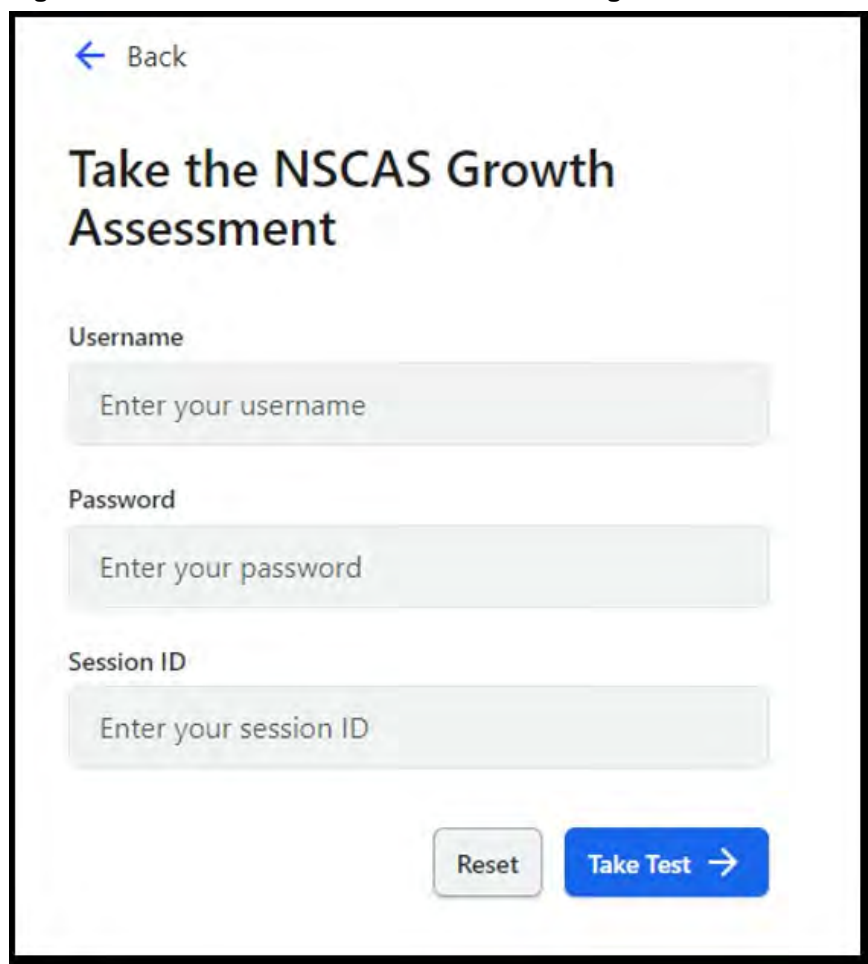
Content Area	Grade	#FT Items in the Pool	#FT Items Administered	Data Review				#Total Accepted Items
				#Included	#Accepted	#Rejected /DNU	#Revise /ReFT	
ELA	3	166	161	23	146	7	8	146
	4	133	131	39	100	12	19	100
	5	171	170	46	136	18	16	136
	6	144	141	38	116	19	6	116
	7	155	150	49	120	15	15	120
	8	191	191	56	145	17	29	145
Mathematics	3	13	13	12	9	3	1	9
	4	3	3	2	2	1	0	2
	5	6	6	6	3	3	0	3
	6	32	32	24	16	8	8	16
	7	10	10	9	7	2	1	7
	8	5	5	4	2	2	1	2
Science	5	119	119	24	106	0	13	106
	8	134	134	32	128	0	6	128

Section 3: Test Administration and Security

The Spring 2023 NSCAS testing window was from April 3–May 5, 2023, and the make-up testing window was from May 8–12, 2023. The tests were untimed and administered online via the NSCAS Growth Platform. Testing sessions were structured as a single session, although students could complete the tests in more than one sitting by pausing the test. Students were not able to go back to previous items.

The NSCAS Growth Platform test management system is a roles-based platform that allows users to roster students, set up test sessions, and administer the assessment. Figure 3.1 presents the student NSCAS Growth Platform login screen. NSCAS Growth Platform works with the NWEA secure lockdown testing browser to administer the assessments, which is required for NSCAS testing. Paper-pencil versions were also available as an accommodation. Each district was required to return either a paper-pencil answer sheet or an online record for all grades 3–8 students enrolled in the district.

Figure 3.1. NSCAS Growth Platform Student Login Screen



The NSCAS administration supported student testing on Windows® PC, Macintosh®, iPads, and Chromebooks that met the following specifications. Touch screens were not supported, and Chromebook tablets were only supported if the student was using an external keyboard. iPad

mini® devices were not recommended. The *NSCAS System and Technology Guide* has system requirements (p. 6).⁵

3.1. User Roles and Responsibilities

Table 3.1 summarizes the user roles and responsibilities for the NSCAS test administration.

Table 3.1. User Roles and Responsibilities

User	Roles and Responsibilities
District Assessment Coordinator	Responsible for coordinating the testing activities of all schools within their districts. Responsibilities include but are not limited to coordinating the test schedules of the schools within the district and setting up test sessions.
School Assessment Coordinator	Serve as single points of contact at the schools for the District Assessment Coordinators and are responsible for coordinating the testing activities within their schools. Responsibilities include but are not limited to secure handling of test materials, such as test tickets, and coordination of proctors. A School Assessment Coordinator and District Assessment Coordinator might be the same person depending on the district's decisions.
Proctor	Responsible for administering the tests to students.

District Assessment Coordinators are responsible for scheduling the test for all schools within the district and coordinating the distribution and collection of test materials, as well as any specific training that the district feels is needed. It is recommended that District Assessment Coordinators conduct an orientation session for School Assessment Coordinators to review and/or discuss:

- District test schedule
- General information in the *NSCAS Growth Assessment Coordinator Guide*
- Procedures for distribution and collection of test materials
- Procedures for maintaining security, as outlined in the *NSCAS Growth Assessment Coordinator Guide* and the *NSCAS Security Manual*
- Proctor orientation

School Assessment Coordinators are responsible for providing secure test materials to proctors and conducting proctor orientations, reviewing topics such as:

- Test schedule
- Administration preparation
- Students with special needs
- Testing conditions
- Security

3.2. Administration Training

In addition to district- and school-held training, NWEA (in collaboration with NDE) held five trainings for district leaders in advance of testing. The Fall 2022 regional workshops were a half-day, virtual workshop held across multiple regions of the state. Information on the spring administration (including test sessions, accessibility, and student rostering) was presented. The

⁵ <https://cdn.nwea.org/docs/NE/SystemTechnologyGuide.pdf>

test administration workshops were two-hour virtual sessions that provided important information on the NSCAS assessments.

3.3. Item Type Samplers

Item Type Samplers are available online and in PDF paper-pencil formats for all content areas and grades and are available on the NSCAS Assessment Portal at https://nwea.force.com/nweaconnection/s/nebraska-practice-tests?language=en_US. The username and password for the item samplers are available in the *Item Type Sampler Manual* (username = ne, password = sampler). Large print and Braille versions were also created and available for order.

The Item Type Samplers are not adaptive. For ELA and mathematics, the Item Type Sampler has 20 items for each respective grade in a content area. The science Item Type Sampler has 12 questions for grade 5 and 13 questions for grade 8. They are also untimed, although the estimated test-taking time for each is 40 minutes. Unlike the actual assessments, progress on an item sampler is not saved; if a student does not complete the test in one sitting, they have to take the entire test again if they restart it. A score is not generated at the end of the test, but keys are made available.

The *Item Type Sampler Manual* is provided on the NSCAS Assessment Portal with information on the item samplers, how to access them, and recommended proctor scripts. The purpose of the item samplers is to allow students to experience the types of items, tools (e.g., calculator), and item aids (e.g., highlighter) available on the actual assessments. They also allow other stakeholders (such as parents and administrators) to experience the assessment environment. For the best student experience, it is recommended that students view the Online Student Tutorial located on the NSCAS Assessment Portal to learn about the available tools and their uses before taking the item samplers. Text-to-speech is available for all item sampler tests, but it is recommended that it only be enabled for students with a documented need on an Individualized Education Plan (IEP) or 504 Plan to be consistent with the requirements for use on the NSCAS assessment.

3.4. Accommodations and Accessibility Features

Table 3.2.2 presents the accessibility supports available for the Spring 2023 NSCAS test administration, including the embedded and non-embedded accommodations and universal features. More information and guidance about these supports can be found in the *NSCAS Accessibility Manual* (NDE, 2023).

- Accommodations are changes in procedures or materials that ensure equitable access to instructional and assessment content and generate valid assessment results for students who need them. Embedded accommodations (e.g., text-to-speech) are provided digitally through instructional or assessment technology, while non-embedded accommodations (e.g., computation supports) are provided locally. Accommodations are available for students for whom there is a documented need on an IEP or 504 Plan.
- Universal features are accessibility supports that are embedded and provided digitally through instructional or assessment technology (e.g., answer choice eliminator) or non-embedded and provided non-digitally at the local level (e.g., scratch paper). Universal features are available to all students as they access instructional or assessment content.

Supports, such as linguistic supports and aids for English language learners (ELLs), were also available to students, either universally or according to need (i.e., IEP or 504 Plan). A complete list of linguistic supports is included in the *NSCAS Accessibility Manual* (NDE, 2023).

Table 3.2. Accommodations and Universal Features

Support	Description
Embedded Accommodations	
Text-to-speech (TTS) ^a	The student uses this feature to hear generated audio of directions, content, and test items. ELA passages may not be read aloud.
Embedded Calculator for all items ^a	The student's disability affects math calculation but not reasoning.
Non-Embedded Accommodations	
Paper-pencil	The student takes the assessment on paper instead of online.
Math supports	For students who need additional supports for math (e.g., abacus, calculation device, number line, addition/multiplication charts, etc.)
Assistive technology	Includes such supports as typing on customized keyboards, assistance with using a mouse, mouth, head stick, or other pointing devices, sticky keys, touch screen, trackball, speech-to-text conversion, or voice recognition
Audio amplification device	A hearing-impaired student uses an amplification device (e.g., FM system, audio trainer)
Braille	A raised-dot code that individuals read with the fingertips. Graphic material is presented in a raised format.
Braille writer or notetaker	A blind student uses a braille writer or note-taker with the grammar checker, internet, and file-storing functions turned off.
Flexible scheduling	The number of items per session can be flexibly defined based on the student's need.
Large print test booklet	A large print form of the test is provided to a student with a visual impairment. A student may respond directly in the test booklet, and a test administrator transfers answers onto an answer document.
Project online test	An online test is projected onto a large screen or wall. The student must use an alternate supervised location that does not allow others to view test content.
Primary mode of communication	The student uses communication device, pointing, or other mode of communication to communicate answers.
Read aloud	Only for students who have a documented need for paper-pencil. The student will have those parts of the test that have audio support in the computer-based version read by a qualified human reader in English.
Response assistance	The student responds directly in the test booklet and a test administrator transfers answers onto an answer sheet.
Scribe	The student dictates their responses to an experienced educator who records verbatim what the student dictates.
Sign interpretation	An educational sign language interpreter signs the test directions, content, and test items to the student. ELA passages may not be signed. The student may also dictate responses by signing.
Specialized presentation of test	Examples include colored paper, tactile graphics, color overlay, magnification device, and color of background.

Support	Description
Embedded Universal Features	
Answer choice eliminator ^a	Used to cross out answer choices that do not appear to be correct.
Color contrast ^a	Background color can be adjusted based on the student's need.
Highlighter ^a	Used for marking desired text, items, or response options with a color.
Keyboard navigation	The student can navigate throughout test content by using a keyboard (e.g., arrow keys). This feature may differ depending on the testing platform or device.
Line reader/line guide	Used as a guide when reading text.
Math tools ^a	These digital tools (e.g., ruler, protractor, calculator) are used for tasks related to math items. They are available only with the specific items for which one or more of these tools would be appropriate.
Notepad ^a	Used as virtual scratch paper to make notes or record responses.
Zoom (item-level)	The student can enlarge the size of text and graphics on a given screen. This feature allows students to view material in magnified form on an as-needed basis. The student may enlarge test content at least fourfold. The system allows magnifying features to work in conjunction with other accessibility features and accommodations provided.
Non-Embedded Universal Features	
Alternate location	The student takes the test at home or in a care facility (e.g., hospital) with direct supervision. For facilities without internet, a paper-pencil test will be allowed.
Directions	The test administrator rereads, simplifies, or clarifies directions aloud for the student as needed.
Flexible scheduling	Districts and schools have flexibility to schedule each content test. Each test is only a single session and can be scheduled for one or multiple days.
Cultural considerations	The student receives a paper-pencil form due to a specific belief or practice that objects to the use of technology. This student does not use technology for any instructional-related activities. Districts must contact NDE to request this accessibility feature.
Noise buffer/headphones	The student uses noise buffers to minimize distraction or filter external noise during testing.
Redirection	The test administrator directs/redirects the student's focus on the test as needed.
Scratch paper (plain or graph)	The student uses blank scratch paper, blank graph paper, or an individual erasable whiteboard to make notes or record responses.
Setting	The student is provided a distraction-free space or alternate, supervised location (e.g., study carrel, front of classroom, alternate room).
Student reads test aloud	The student quietly reads the test content aloud to themselves. This feature must be administered in a setting that is not distracting to other students.
Medical device	The student may have access to an electronic device for medical purposes (e.g., glucose monitor).
Focus/Engagement assistance	The student may have access to items/conditions (e.g., fidgets, flexible seating, water bottle at student's desk, music for individual students with headphones,

Support	Description
	gum/mints) they typically have access to during regular instruction to help focus and/or engagement.

^a Not available for NSCAS Alternative Assessments

3.5. User Acceptance Testing (UAT)

User acceptance testing (UAT) is conducted each term to test the most common configurations in use in Nebraska on each device based on the following criteria:

- Content
- Item type functionality (e.g., make sure the correct answer can be selected for a multiple-choice item)
- Universal features/item aids and tools (e.g., highlighter, eraser, answer eliminator)
- Item-specific features (e.g., ruler, protractor)
- Accessibility features (e.g., TTS)
- New features/enhancements

Testers are NWEA staff who are familiar with how the functionality is supposed to work. In addition to a training and kick-off on the process and a checklist of tasks, technical product managers are present at the kick-off meeting to describe the UAT process overall, expected enhancements to functionality, and known issues. Use cases describing each item feature and other support documentation are provided to testers to review prior to UAT. Testers spend 1–2 hours reviewing existing documentation prior to performing testing. They also explore the Item Type Sampler beforehand.

To conduct UAT, testers are assigned tests on a particular device and location (e.g., work desk, at home) and spend approximately 30–40 minutes per test. Bugs are reported and tracked manually. Triage meetings take place to review all new reported entries and to update the status for known issues. During the UAT process, testers review live, secure NSCAS tests. Test security is taken very seriously, and testers are not allowed to share, copy, record, or take photos of the items they review.

NWEA staff review the data produced from UAT to ensure they conform to expectations for completed tests, tests assigned not-tested codes (NTCs), incomplete tests, tests that were reset, and additional activities that occur during testing. User roles are tested for accessibility and functionality. Operational and score reports are reviewed to ensure they meet requirements.

3.6. Student Participation

All students with disabilities were expected to participate in the NSCAS. No student, including students with disabilities or who require a paper assessment, can be excluded from the state assessment and accountability system. All students are required to have access to grade-level content, instruction, and assessment. Students with disabilities may be included in state assessment and accountability in the following ways:

- Students are tested on the NSCAS without accommodations.
- Students are tested on the NSCAS with approved accommodations specified in the student's IEP. Accommodations provided to students must be specified in the student's IEP and have been used during instruction throughout the year.

- Students can be tested with the NSCAS Alternate assessment if they qualify. Only students with the most significant cognitive disabilities (typically less than 1% of students) can take these tests. The NSCAS Alternate assessment is distributed and administered by DRC.

Use of non-approved accommodations may invalidate the student's score. Non-approved accommodations used in state testing result in both a zero score and no participation credit. Accommodations provide adjustments and adaptations to the testing process that do not change the expectation, grade level, construct, or content being measured. Accommodations should only be used if they are appropriate for the student and have been used during instruction throughout the year. In contrast, modifications are adjustments or changes in the test that affect test expectations, grade level, construct, or content being measured. Modifications are not acceptable in the NSCAS assessments.

3.6.1. *Paper-Pencil Participation Criteria*

Students participating in the paper-pencil administration have to meet one of the following criteria:

- The student has a medical condition that does not allow the use of computer screens.
- The student requires Braille/large print.
- The facility does not allow internet access.
- The student requires written translations of languages other than Spanish.
- Cultural considerations must be taken into account.
- The student needs the test in both English and another language side-by-side (mathematics and science only).
- The student is an English language learner with limited prior access to technology.

3.6.2. *Participation of English Language Learners (ELLs)*

According to the Elementary and Secondary Education Act (ESEA), ELLs are students who have a native language other than English, OR who came from an environment where a language other than English has had a significant impact on their level of English proficiency, AND whose difficulties in speaking, reading, writing, or understanding the English language may be sufficient to deny the individual (i) the ability to meet the state's proficient level of achievement on state assessments, (ii) the ability to successfully achieve in classrooms where the language of instruction is English, or (iii) the opportunity to participate fully in society (NCLB, 2002).

Each district with ELL students should have a written operational definition used for determining services and meeting Office of Civil Rights requirements. Both state and federal laws require the inclusion of all students in the state testing process. ELL students must be tested on the NSCAS assessments. Districts should have reviewed the following guidelines before testing:

- In determining appropriate linguistic supports for students in the NSCAS system, districts should use the *NSCAS Accessibility Manual* (NDE, 2023).
- Districts must be aware of the difference between linguistic supports (accommodations for ELLs) and modifications.
- For students learning the English language, linguistic supports are changes to testing procedures, testing materials, or the testing situation that allow the students meaningful

participation in the assessment. Effective linguistic supports for ELL students address their unique linguistic and socio-cultural needs. Linguistic supports for ELL students may be determined appropriate without prior use during instruction throughout the year.

- Modifications are adjustments or changes in the test or testing process that change the test expectation, grade level, construct, or content being measured. Modifications are not acceptable in the NSCAS assessments.

3.6.3. Participation of Recently Arrived Limited English Proficient Students

Recently Arrived Limited English Proficient (RAEL) students are defined by the U.S. Department of Education as students with limited English proficiency who have attended schools in the United States for fewer than 12 months. The phrase “schools in the United States” includes only schools in the 50 states and the District of Columbia; it does NOT include Puerto Rico. Districts must assess all RAEL students on all NSCAS assessments each year based on the grade level of the student using linguistic supports.

3.7. Test Security

In a centralized testing process, it is critical that equity of opportunity, standardization of procedures, and fairness to students is maintained. Therefore, NDE asks that all school districts review the NSCAS security procedures provided in the *NSCAS Growth Assessment Coordinator Guide*. Breaches in security are taken very seriously, and it is emphasized that they must be quickly identified and reported to NDE’s Statewide Assessment Office. Districts are encouraged to maintain a set of policies that includes a reference to Nebraska’s *NSCAS Security Manual*. A sample district testing and security policy is included in Nebraska’s Standards, Assessment, and Accountability Updates posted on NDE’s website. Whether districts use this sample, the procedures offered by the State School Boards Association, or policies drafted by other law firms, local district policy should address the *NSCAS Security Manual*. NDE encourages all districts with questions to contact their own local school attorney for customization of such a policy.

As part of NDE’s security policy, the principal of each school participating in the NSCAS assessments is required to complete and sign a Building Principal Security Agreement and return it to the Statewide Assessment Office. District Assessment Coordinators are required to complete and sign the District Assessment Coordinator Confidentiality of Information Agreement and return it to the Statewide Assessment Office. School districts are bound to hold all certificated staff members in school districts accountable for following the Regulations and Standards for Professional Practice Criteria as outlined in Rule 27. The *NSCAS Security Manual* is intended to outline clear practices for appropriate security.

3.7.1. Test Security

3.7.1.1. Physical Warehouse Security

All NWEA personnel—including subcontractors, vendors, and temporary workers who have access to secure test materials—are required to agree to keep the test materials secure and sign security forms that state understanding of the secure nature of test items and the confidentiality of student information. Access to the NWEA headquarters is by badged-security access. All visitors entering the facility are required to sign in at the front desk and obtain an entry badge that allows them access to the facility. The following additional security procedures are maintained for the NSCAS program:

- Test materials received from the printing subcontractors are stored in a room at NWEA headquarters prior to packaging and shipping to districts.

3.7.1.2. Secure Destruction of Test Materials

Printed materials for the Spring 2023 administration are not considered secure, therefore districts are authorized to destroy material locally.

3.7.1.3. Shipping Security

For district shipments, NWEA uses secure and trackable UPS ground and two-day shipping services to send materials to and receive materials from districts. The system interfaces with the in-house UPS shipping system, thus making certain that deliveries are made to accurate and correct addresses. Address verification is used to ensure that the materials are shipped to known UPS addresses before shipping. Every box is assigned a unique UPS tracking number.

3.7.1.4. Electronic Security of Test Materials and Data

All computer systems that store test materials, test results, and other secure files require password access. During the test-material printing processes, electronic files are transferred via a server accessed by Secure File Transfer Protocol (SFTP). Access to the site is password controlled and on an as-needed basis. Transmission to and from the site is via an encrypted protocol. Transfer of student data between NWEA and print vendors follows secure procedures. Data files are exchanged through an SFTP site and the secure application program interface.

3.8. Partner Support

The NWEA Partner Support Services team provided implementation and technical support throughout the 2022–2023 school year for the NSCAS assessments. This team provides resources to support Nebraska and its educators, assisting with generating roster files, configuration of the assessment program, accessing online reports, and general questions about the use of the online assessment system. NWEA provides phone, email, and chat support to schools and educators from 8:00 a.m. to 5:00 p.m. Central Time (CT) Monday through Friday, and 7:00 a.m. to 5:00 p.m. CT during the testing windows, as described in Table 3.3. Table 3.4 presents the number of cases presented to the Partner Support team by case type for the entire 2022–2023 school year for the NSCAS tests. More than half of the cases were related to testing (i.e., administration questions).

Table 3.3. Partner Support Communication Options

Phone Support	NWEA uses Voice Over Internet Protocol (VOIP) phone systems to allow callers to quickly reach the first available representative. VOIP also provides remote-access capabilities for our staff, enabling Partner Support team members to provide seamless service even during times of inclement weather or office closure. Reports from our phone system and customer-relationship management tool, as well as call-monitoring tools, are used in monitoring quality and in determining additional training needs.
Email Support	Emailed support requests are also handled quickly and efficiently. It is NWEA's goal to respond to all emails within 24 hours from time of receipt. Emails received within NWEA business hours are responded to on the same business day.
Chat Support	Chat is a convenient method of contacting support for in-the-moment questions or for use in the rare occurrence of a phone-service disruption.

Table 3.4. Number of NSCAS Cases to Partner Support in 2022–2023

Case Type	#Cases	% of Total Cases
Student Mobility	14	< 2%
Reports	130	15%
Navigation	54	6%
Setup and Management	332	38%
Testing	336	39%
Total	866	100%

NWEA monitors all service activities through daily, weekly, and monthly reports and makes adjustments as needed to ensure appropriate coverage for Nebraska support needs during peak use times, such as prior to and throughout the testing windows. All Tier 1 and Tier 2 support staff members are required at hire to undergo a two-week training program led by the NWEA Senior Support Specialist team and team trainers. The training program consists of a combination of instructor-led and self-paced eLearning courses, covering all relevant team policies and procedures, including security requirements for handling student data, product expertise, and troubleshooting requirements. In addition, several days of “phone shadowing” are built into the program to ensure that each new staff member has the opportunity to participate in calls with veteran staff monitoring prior to working independently. Senior Support Specialists are responsible for continually updating training program content to ensure that all support team staff members are knowledgeable of current policies. In addition, project managers and product training resources are dedicated to NDE’s program to train support staff on Nebraska-specific policies. On average, each state team member participates in four hours of training related to Nebraska programs.

Section 4: Scoring and Reporting

The online ELA and mathematics assessments were administered adaptively via NWEA's constraint-based engine (Cadabra), whereas the science assessments were administered as a fixed form. For science, each grade had 20 different forms, but the operational items were the same across all forms. Also, all paper-pencil tests and all Spanish versions were administered as a fixed form.

4.1. Scoring Rules

An attemptedness rule is the minimum number of items a student must attempt during testing to be included in psychometric analyses and/or receive a numeric score. Table 4.1 presents the attemptedness rules for scoring.

Table 4.1. Attemptedness Rules for Scoring

#OP Items Attempted	Include in Psychometric Analyses?	Receive Scale Score?	Receive Achievement Level?
0	No	Yes, LOSS	Yes, lowest level
1–9	No	Yes, LOSS +1	Yes, lowest level
10+	Yes	Yes, calculated MLE scores	Yes

Note. LOSS = lowest obtainable scale score; MLE = maximum likelihood estimation

The attemptedness rule was decided based on the results of the standard error of measurement (SEM) that became relatively stable (i.e., SEM became less than 1.0 for students in the middle of true theta distribution) after 10 operational items from the simulation data and the finding of a small number of 2017 students who attempted less than 10 items. Regarding scoring, NWEA ran analyses using a subpopulation of the 2017 students and found that the number of not-reached items increased the amount of estimation error, suggesting larger estimation error with the penalty function (i.e., to score those not-reached items as wrong). However, scoring consistency was also considered for fixed forms (science). Thus, NDE made the following scoring rules in consultation with the State and District Coordinators, as summarized in Table 4.2:

1. Students who took the adaptive assessment (i.e., the ELA and mathematics online adaptive forms) received straight maximum likelihood estimation (MLE) scoring (i.e., regular MLE scoring with no penalty) regardless of test-completion status. Students who took the Spanish online assessment also received straight MLE scoring.
2. Except for the Spanish online form, MLE scoring with penalty was applied to fixed forms (i.e., science online and paper-pencil, Spanish paper-pencil, and ELA and mathematics paper-pencil), treating omit and multi-marks as incorrect.
3. Sub-scores were provided for students who attempted a minimum of 10 items overall and 4 items within each specific reporting category.

Table 4.2. MLE Scoring

Content Area	English Form		Spanish Form		Breach Form
	Online	Paper-Pencil	Online	Paper-Pencil	Paper-Pencil
ELA/Mathematics	No penalty	With penalty	No penalty	With penalty	With penalty
Science	With penalty	With penalty	With penalty	With penalty	With penalty

4.2. Score Reporting Methods

Student performance on the NSCAS assessment is reported as a scale score and achievement level. Each content area is scaled separately. Therefore, the scale scores for one content area cannot be compared with another content area. For ELA and mathematics, NSCAS Growth reports also provide estimated RIT scores for students who complete the test. Table 4.3 presents the score ranges for both scores.

Table 4.3. Score Range (LOSS and HOSS) for NSCAS Scale Score and Estimated RIT Score

Content Area	Grade	NSCAS Scale Score			Estimated RIT Score	
		LOSS	HOSS	Calculated LOSS ^a	LOSS	HOSS
ELA	3	2220	2840	2222	100	350
	4	2250	2850	2252	100	350
	5	2280	2860	2282	100	350
	6	2290	2870	2292	100	350
	7	2300	2880	2302	100	350
	8	2310	2890	2312	100	350
Mathematics	3	1000	1470	1002	100	350
	4	1010	1500	1012	100	350
	5	1020	1510	1022	100	350
	6	1030	1530	1032	100	350
	7	1040	1540	1042	100	350
	8	1050	1550	1052	100	350
Science	5	3000	3250	3002	–	–
	8	3000	3250	3002	–	--

^a Calculated LOSS = lowest calculated score for students with 10 or more OP items attempted.

An achievement level is a written description of the student's overall performance and is used to help make the scale scores meaningful. There are three other important reasons for establishing achievement levels:

- To give meaning to the scale scores in order to help Nebraska students and parents use the results effectively
- To connect the scale scores on the tests to the content standards in order to assist Nebraska educators in supporting students to become college and career ready
- To meet the requirements of the U.S. Department of Education

The Nebraska State Board of Education defines three achievement levels for each content area, as shown in Table 4.4.

Table 4.4. Achievement Level Descriptions for ELA, Mathematics and Science

Achievement Level	Description
<i>Developing</i>	<i>Developing</i> learners <u>do not yet demonstrate proficiency</u> in the knowledge and skills necessary at this grade level, as specified in the assessed Nebraska College and Career Ready Standards. These results provide evidence that the

Achievement Level	Description
	student may need additional support for academic success at the next grade level.
<i>On Track</i>	<i>On Track</i> learners demonstrate proficiency in the knowledge and skills necessary at this grade level, as specified in the assessed Nebraska College and Career Ready Standards. These results provide evidence that the student will likely be ready for academic success at the next grade level.
<i>Advanced</i>	<i>Advanced</i> Benchmark learners demonstrate advanced proficiency in the knowledge and skills necessary at this grade level, as specified in the assessed Nebraska College and Career Ready Standards. These results provide evidence that the student will likely be ready for academic success at the next grade level.

The reporting categories in Table 4.5 were to be used for scoring and reporting. Items were mapped to a reporting category based on the indicators. For science, reporting category scores were not provided in 2023.

Table 4.5. Reporting Categories

Content Area	Reporting Category
ELA (Fall)	Reading Vocabulary Reading Comprehension Vocabulary Writing Skills
ELA	Reading Prose and Poetry Reading Informational Text Vocabulary Writing and Foundations of Writing
Mathematics	Number Algebra Geometry Data

Note. New standards and reporting categories in ELA were implemented beginning with Winter 2022–2023.

4.3. Report Summary

The following reports were prepared for the 2023 NSCAS test administration. Examples of the reports and additional information can be found in the Interpretive Guide.⁶

- State Level
 - Student Score Data File
 - Organization Report—State level
 - State Demographic Report
- Region
 - Organization Report—Region level
 - Region Demographic Report
 - Region Roster

⁶ https://www.nwea.org/uploads/NSCASReportsIntGuideEnglish_NWEA_Guide.pdf

- District Level
 - Student Score Data File
 - Organization Report—District level
 - District Demographic Report
 - District Roster
- School Level
 - Organization Report—School level
 - School Roster
 - School Demographic Report
- Class/Group Level
 - Class/Group Roster
- Student Level
 - Dynamic Student Report
 - Student Growth Report
 - Individual Student Report (ISR)— English
 - Individual Student Report (ISR)— Spanish (Spring only)

ISRs show a student's performance on the NSCAS Growth tests. Content areas are combined to produce a single ISR report for a student. ISRs are available through the NSCAS Growth platform and shipped to the districts. Some ISRs are shipped to their new fall enrollment district, while others are shipped to their reportable district. If a not-tested code (NTC) is applied to a content area, the student's achievement level scores are reported as affected by the NTC, as defined in Table 4.6. If a student has an NTC of INV, PAR, STR, or UTT assigned to their test, the automatically assigned score displays with a score of the lowest obtainable scale score (LOSS) for that grade and content area.

Table 4.6. Not-Tested Codes (NTCs)

Code	Name/Description	Included in Reports	Scoring
ALT	Alternate Assessment: Student took the NSCAS Alternate assessment and is not included in results from this testing vendor.	FALSE	No score provided
EMW	Emergency Medical Waiver: Student was not tested because of an approved Emergency Medical Waiver.	TRUE	No score provided
EXP	Exempt: Student exempt from testing due to certain circumstances, such as student requiring unavailable accommodation; student is attending an out-of-state facility; or testing irregularities.	FALSE	Score not included in reports or calculations
FTE	Full-Time Equivalency: Full-time equivalency is less than 51%, so the student is excluded from testing.	FALSE	Score not included in reports or calculations
INV	Invalid: Student's assessment was invalidated, such as for a security breach or student refuses to finish the test.	TRUE	Score as LOSS
LBW	Left Before Window: Student withdrew from the district or school before the test	FALSE	Score not included in reports or calculations

Code	Name/Description	Included in Reports	Scoring
	window began. Excluded from reporting. ADVISER enrollment data must support coding.		
NCE	Not Currently Enrolled: Student was not enrolled in the district/school during testing window.	FALSE	Score not included in reports or calculations
OTH	Other: Student was not tested for reasons not covered by other descriptions. For example, occurrence of a natural disaster.	TRUE	Score suppressed
PAR	Parent Refusal: Student was not tested because of a formal request from parent or guardian.	TRUE	Score as LOSS
RMV	Remove: Student left the district before the test window; student is a full-time home-schooled student; or there are duplicate student records.	FALSE	Score not included in reports or calculations
STR	Student Refusal: Student was not tested due to student refusal to participate.	TRUE	Score as LOSS
UTT	Unable to Test: District was unable to test the student during the testing windows due to excessive absences or suspension/expulsion.	TRUE	Score as LOSS

4.3.1. Report Verification

The NSCAS report quality assurance (QA) process consists of validating the data and reports using the scoring specifications, reporting specifications, mockups, layouts, scale scores, and cut information.

The objectives of report verification are to ensure that:

- The reports match NDE's expectations.
- The data on the report are accurate.
- The data on the report are presented per NDE's expectations.
- NDE and users can access the reports.

The following report segments are checked during the QA process:

- Formatting
- Static text (text that does not change)
- Dynamic text (text that changes)
- Student data (demographic information)
- Score-related data (scale scores, achievement levels)
- Historical charts and data footnotes
- NTC behavior
- Not enough items (NEI) behavior
- Sorting (sort order of the report)

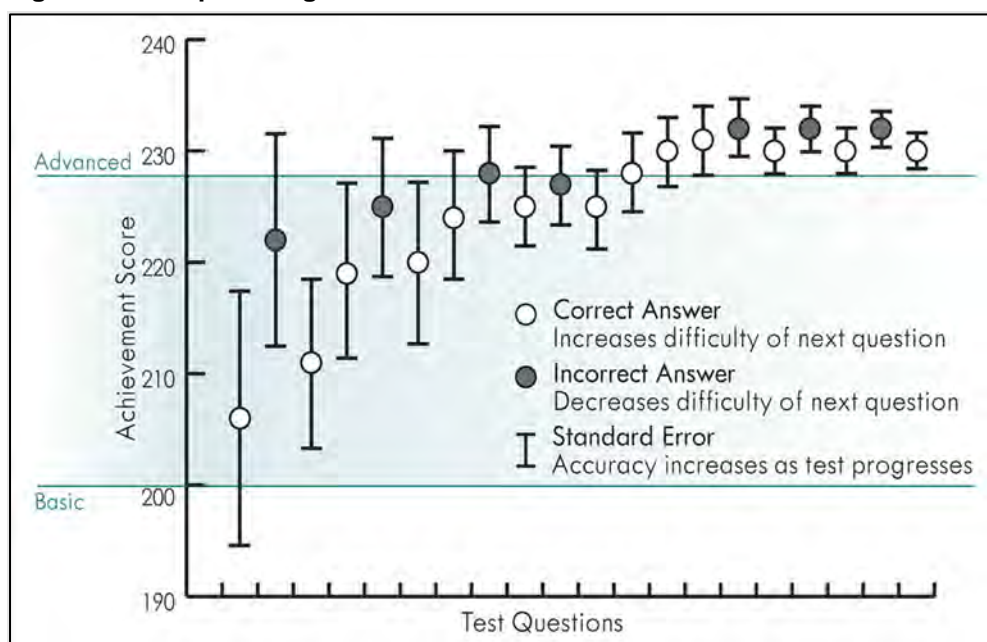
- Naming conventions for reports, files, and folders
- Similar data is the same across all reports
- Summation of data
- User interface functionality

Section 5: Adaptive Engine

5.1. Overview

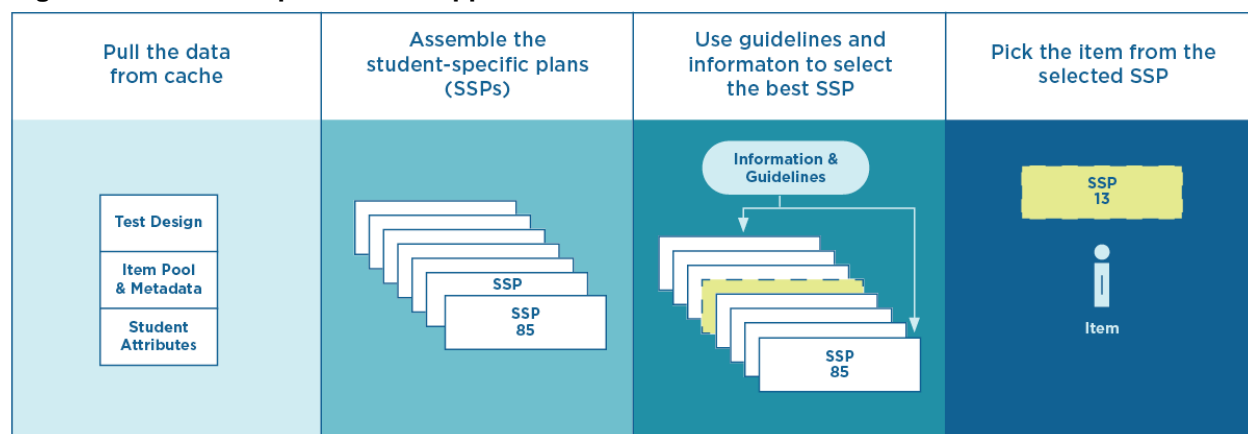
A computer adaptive test (CAT) administers items during assessment to match the ability level of the student. Students receive different items based on item difficulty and their ability levels. For example, students with lower ability levels (based on their answers to previous items) receive easier items compared with students with higher ability levels, who receive harder items as the test progresses. The adaptive engine of NWEA, Cadabra, uses the table of specifications (TOS) and a student's momentary theta (θ) to drive item selection, as shown in Figure 5.1. Momentary theta is the ability estimate of the student that is recalculated and updated after each item is answered.

Figure 5.1. Adaptive Engine Overview



Items are selected based on item difficulty. The goal of the adaptive constraint-based engine's item selection is to provide a test that meets "must-have" constraints and "nice-to-have" guidelines. Cadabra has two stages of consideration as it selects the items necessary to conform to the test blueprint while providing the maximum information about the student based on the student's momentary ability estimate. The student-specific plan (SSP), similar to the shadow test approach (Van der Linden & Reese, 1998), selects items based on the required aspects of the test blueprint and the student's momentary theta, as shown in Figure 5.2. Item selection for the SSP occurs through a process of choosing multiple feasible SSPs and then choosing the complete SSP that best maximizes guideline adherence and information. Only after the best SSP has been chosen are items ordered (NWEA, 2020).

Figure 5.2. Student-Specific Plan Approach



Note. Selections are based on the similar shadow test approach.

The following updates were made for Fall 2022–2023:

- The operational test in Fall 2022–2023 has a total of 40 items, just like the Winter 2021–2022 forms.
- On-grade diagnostic items (i.e., MAP Growth items) are allowed in the operational (i.e., accountability) section as well.
- Diagnostic items are allowed up to two grades above and unlimited grades below, while only adjacent grades were included in 2021–2022 tests.
- Stand-alone reading items from MAP Growth are allowed.
- Blueprint requirements for polytomous items have been loosened.
- Recently field-tested items are reserved for operational use in the spring.
- Student true thetas in simulated student files have been created based on the recalibrated item parameters.
- The fall and winter test models in mathematics have been redesigned for more adaptivity (to be more similar to MAP Growth in that regard), and the summative blueprint is no longer strictly enforced. Therefore, additional flexibility for mathematics does not guarantee that all students will satisfy the 27-item summative blueprint.

The following updates were made for Winter 2022–2023:

- For ELA, new standards and reporting categories have been implemented, starting from Winter 2022–2023.

The following updates were made for Spring 2022–2023:

- The operational test is slightly longer for spring than for fall/winter, just like previous years, having a total of 45 items, including field-test items, while the winter test had a total of 40 items.

5.2. Engine Simulations and Evaluation

Pre-administration engine simulations and post-administration engine evaluation studies are important evidence, along with post-administration analyses, for confirming interpretation and test-score use arguments regarding student proficiency with the state standards.

Pre-administration simulations were conducted prior to the operational testing window to evaluate the engine's item-selection algorithm and estimation of student ability based on the TOS. The simulation tool used the operational engine, thereby providing results with the same properties and functionality as would be seen operationally. Detailed information regarding the simulation study can be found in the full reports (NWEA, 2021a, 2022a).

After the testing window closed, a post-administration evaluation study was conducted to determine whether the adaptive engine performed as expected. Detailed information regarding all results of the post-administration evaluation study can be found in the full reports (NWEA, 2022b, 2022c).

Overall, the engine performed as it should based on the blueprint (i.e., the TOS) constraints. The reporting category points had a 100% match. The adaptive engine also showed a similar performance when estimating students' ability in terms of standard error of measurement (SEM) and reliability. Item exposure rates were also acceptable given that the adaptive engine used almost all items to administer the test, and most used items had a 0–20% exposure rate.

5.2.1. Evaluation Criteria

Computational details of the precision ability estimation statistics (i.e., bias, p value, and MSE) are as follows (CRESST, 2015):

$$\begin{aligned} bias &= N^{-1} \sum_{i=1}^N (\theta_i - \hat{\theta}_i) \\ MSE &= N^{-1} \sum_{i=1}^N (\theta_i - \hat{\theta}_i)^2 \end{aligned}$$

where θ_i is the true score, and $\hat{\theta}_i$ is the estimated (observed) score. To calculate the variance of theta bias, the first-order Taylor series of the above equation is used as follows:

$$var(bias) = \sigma^2 * g'(\hat{\theta}_i)^2 = \frac{1}{N(N-1)} \sum_{i=1}^N (\theta_i - \hat{\theta}_i)^2$$

where $\hat{\theta}_i$ is an average of the estimated theta. Significance of the bias is then tested as follows:

$$Z = bias / \sqrt{var(bias)}$$

A p value for the significance of the bias is reported from this z-test with a two-tailed test. The average standard error (SE) is computed as follows:

$$Mean(se) = \sqrt{N^{-1} \sum_{i=1}^N se(\hat{\theta}_i)^2}$$

where $se(\hat{\theta}_i)^2$ is the standard error of the estimated θ for individual i . To determine the number of students falling outside the 95% and 99% confidence interval coverage, a t -test was performed as follows:

$$t = \frac{\theta_i - \hat{\theta}_i}{se(\hat{\theta}_i)}$$

where $\hat{\theta}_i$ is the ability estimate for individual i , and θ_i is the true score for individual i . The percentage of students' estimated theta falling outside the coverage was determined by comparing the absolute value of the t -statistic with a critical value of 1.96 for 95% coverage and 2.58 for 99% coverage.

Traditional reliability coefficients from classical test theory consider individual items and depend on all test takers to take common items; however, in a CAT, different students receive different items. Therefore, NWEA calculated the marginal reliability coefficient for the CAT administration. Samajima (1994) recommends the marginal reliability coefficient because it uses test information (e.g., variance of estimated theta and SEM) to estimate the reliability of student scores:

$$\text{Marginal Reliability} = \frac{\text{var}(\hat{\theta}) - \sigma^2}{\text{var}(\hat{\theta})}$$

where σ is defined as:

$$\sigma = E\{[I(\theta)]^{-1/2}\}$$

5.2.2. Blueprint Constraint Accuracy

Table 5.1 through Table 5.6 present the blueprint constraint results at the reporting category level for the pre-administration simulation study and the post-administration engine evaluation study for fall, winter, and spring, respectively. For ELA, the number of items at the reporting category level resulted in a 100% match for all grades based on the blueprint, with marginal deviation in the number of points based on the availability and selection of polytomously scored items. Note that new standards and reporting categories have been implemented starting from Winter 2022–2023 in ELA. For mathematics, the fall and winter test models have been redesigned for more adaptivity (to be more similar to MAP Growth in that regard), and the summative blueprint is no longer strictly enforced. Therefore, additional flexibility for mathematics does not guarantee that all students will satisfy the 27-item summative blueprint.

Table 5.1. Blueprint Constraint Accuracy by Reporting Category—Fall Simulations

Grade	Reporting Category	#Items			#Points		
		Min.	Max.	%Match	Min.	Max.	%Match
ELA							
3	Reading Vocabulary	6	6	100	6	7	100
	Reading Comprehension	14	14	100	16	17	98
	Writing Skills	7	7	100	9	9	100
4	Reading Vocabulary	6	6	100	6	7	100
	Reading Comprehension	14	14	100	16	17	98
	Writing Skills	7	7	100	9	9	100
5	Reading Vocabulary	6	6	100	6	7	100
	Reading Comprehension	14	14	100	16	17	98
	Writing Skills	7	7	100	9	9	100
6	Reading Vocabulary	6	6	100	6	7	100
	Reading Comprehension	14	14	100	16	17	100
	Writing Skills	7	7	100	9	9	100
7	Reading Vocabulary	6	6	100	6	7	100
	Reading Comprehension	14	14	100	16	17	99
	Writing Skills	7	7	100	9	9	100
8	Reading Vocabulary	6	6	100	6	7	100
	Reading Comprehension	14	14	100	16	17	99
	Writing Skills	7	7	100	9	9	100
Mathematics							
3	Number	10	10	95	11	12	99
	Algebra	5	5	100	6	7	100
	Geometry	7	7	66	8	9	99
	Data	5	5	92	6	7	91
4	Number	10	10	58	11	12	88
	Algebra	6	6	74	7	8	96
	Geometry	6	6	94	7	8	95
	Data	5	5	84	6	7	95
5	Number	10	10	81	11	12	94
	Algebra	6	6	87	7	8	98
	Geometry	6	6	93	7	8	99
	Data	5	5	98	6	7	100
6	Number	7	7	97	8	9	98
	Algebra	10	10	82	11	12	91
	Geometry	5	5	99	6	7	97
	Data	5	5	44	6	7	98
7	Number	6	6	90	7	8	95
	Algebra	9	9	83	10	11	94
	Geometry	5	5	92	6	7	97
	Data	7	7	95	8	9	94
8	Number	7	7	93	8	9	94
	Algebra	7	7	83	8	9	98
	Geometry	8	8	86	9	10	92
	Data	5	5	99	6	7	99

Table 5.2. Blueprint Constraint Accuracy by Reporting Category—Fall Engine Evaluation

Grade	Reporting Category	#Items			#Points		
		Min.	Max.	%Match	Min.	Max.	%Match
ELA							
3	Reading Vocabulary	6	6	100.0	6	7	100.0
	Reading Comprehension	14	14	100.0	16	17	95.8
	Writing Skills	7	7	100.0	9	9	100.0
4	Reading Vocabulary	6	6	100.0	6	7	100.0
	Reading Comprehension	14	14	100.0	16	17	99.6
	Writing Skills	7	7	100.0	9	9	100.0
5	Reading Vocabulary	6	6	100.0	6	7	100.0
	Reading Comprehension	14	14	100.0	16	17	95.3
	Writing Skills	7	7	100.0	9	9	100.0
6	Reading Vocabulary	6	6	100.0	6	7	100.0
	Reading Comprehension	14	14	100.0	16	17	99.6
	Writing Skills	7	7	100.0	9	9	100.0
7	Reading Vocabulary	6	6	100.0	6	7	100.0
	Reading Comprehension	14	14	100.0	16	17	98.8
	Writing Skills	7	7	100.0	9	9	100.0
8	Reading Vocabulary	6	6	100.0	6	7	100.0
	Reading Comprehension	14	14	100.0	16	17	97.9
	Writing Skills	7	7	100.0	9	9	100.0
Mathematics							
3	Number	10	10	94.9	11	12	99.5
	Algebra	5	5	100.0	6	7	99.1
	Geometry	7	7	57.5	8	9	99.6
	Data	5	5	94.4	6	7	94.4
4	Number	10	10	63.0	11	12	92.8
	Algebra	6	6	73.8	7	8	96.9
	Geometry	6	6	97.9	7	8	98.1
	Data	5	5	88.4	6	7	98.9
5	Number	10	10	84.1	11	12	95.8
	Algebra	6	6	88.0	7	8	98.4
	Geometry	6	6	93.6	7	8	99.5
	Data	5	5	97.5	6	7	86.8
6	Number	7	7	98.5	8	9	97.9
	Algebra	10	10	69.5	11	12	88.9
	Geometry	5	5	99.5	6	7	98.2
	Data	5	5	44.6	6	7	94.5
7	Number	6	6	94.5	7	8	95.4
	Algebra	9	9	74.6	10	11	95.0
	Geometry	5	5	88.3	6	7	98.5
	Data	7	7	97.5	8	9	90.5
8	Number	7	7	95.1	8	9	93.5
	Algebra	7	7	71.3	8	9	98.4
	Geometry	8	8	84.1	9	10	89.4
	Data	5	5	99.5	6	7	99.7

Table 5.3. Blueprint Constraint Accuracy by Reporting Category—Winter Simulations

Grade	Reporting Category	#Items			#Points		
		Min.	Max.	%Match	Min.	Max.	%Match
ELA							
3	Reading Prose and Poetry	7	9	100.0	7	11	100.0
	Reading Informational Text	7	9	100.0	7	11	100.0
	Vocabulary	4	5	100.0	4	7	100.0
	Writing and Foundations of Writing	6	7	100.0	6	9	100.0
4	Reading Prose and Poetry	7	9	100.0	7	11	100.0
	Reading Informational Text	7	9	100.0	7	11	100.0
	Vocabulary	4	5	100.0	4	7	100.0
	Writing and Foundations of Writing	6	7	100.0	6	9	100.0
5	Reading Prose and Poetry	7	9	100.0	7	11	100.0
	Reading Informational Text	7	9	100.0	7	11	100.0
	Vocabulary	4	5	100.0	4	7	100.0
	Writing and Foundations of Writing	6	7	100.0	6	9	100.0
6	Reading Prose and Poetry	7	9	100.0	7	11	100.0
	Reading Informational Text	7	9	100.0	7	11	100.0
	Vocabulary	4	5	100.0	4	7	100.0
	Writing and Foundations of Writing	6	7	100.0	6	9	100.0
7	Reading Prose and Poetry	7	8	100.0	7	10	100.0
	Reading Informational Text	8	9	100.0	8	11	100.0
	Vocabulary	4	5	100.0	4	7	100.0
	Writing and Foundations of Writing	6	7	100.0	6	9	100.0
8	Reading Prose and Poetry	7	8	100.0	7	10	100.0
	Reading Informational Text	8	9	100.0	8	11	100.0
	Vocabulary	4	5	100.0	4	7	100.0
	Writing and Foundations of Writing	6	7	100.0	6	9	100.0
Mathematics							
3	Number	10	10	93.0	11	12	99.0
	Algebra	5	5	100.0	6	7	98.0
	Geometry	7	7	78.0	8	9	100.0
	Data	5	5	96.0	6	7	95.0
4	Number	10	10	67.0	11	12	95.0
	Algebra	6	6	94.0	7	8	97.0
	Geometry	6	6	98.0	7	8	96.0
	Data	5	5	92.0	6	7	96.0
5	Number	10	10	71.0	11	12	95.0
	Algebra	6	6	97.0	7	8	95.0
	Geometry	6	6	80.0	7	8	98.0
	Data	5	5	94.0	6	7	97.0
6	Number	7	7	99.0	8	9	99.0
	Algebra	10	10	78.0	11	12	95.0
	Geometry	5	5	99.0	6	7	97.0
	Data	5	5	84.0	6	7	97.0
7	Number	6	6	90.0	7	8	94.0
	Algebra	9	9	85.0	10	11	95.0
	Geometry	5	5	92.0	6	7	95.0
	Data	7	7	96.0	8	9	99.0
8	Number	7	7	88.0	8	9	90.0
	Algebra	7	7	86.0	8	9	97.0
	Geometry	8	8	86.0	9	10	90.0
	Data	5	5	96.0	6	7	98.0

Table 5.4. Blueprint Constraint Accuracy by Reporting Category—Winter Engine Evaluation

Grade	Reporting Category	#Items			#Points		
		Min.	Max.	%Match	Min.	Max.	%Match
ELA							
3	Reading Prose and Poetry	7	9	100.0	7	11	100.0
	Reading Informational Text	7	9	100.0	7	11	100.0
	Vocabulary	4	5	100.0	4	7	100.0
	Writing and Foundations of Writing	6	7	100.0	6	9	100.0
4	Reading Prose and Poetry	7	9	100.0	7	11	100.0
	Reading Informational Text	7	9	100.0	7	11	100.0
	Vocabulary	4	5	100.0	4	7	100.0
	Writing and Foundations of Writing	6	7	100.0	6	9	100.0
5	Reading Prose and Poetry	7	9	100.0	7	11	100.0
	Reading Informational Text	7	9	100.0	7	11	100.0
	Vocabulary	4	5	100.0	4	7	100.0
	Writing and Foundations of Writing	6	7	100.0	6	9	100.0
6	Reading Prose and Poetry	7	9	100.0	7	11	100.0
	Reading Informational Text	7	9	100.0	7	11	100.0
	Vocabulary	4	5	100.0	4	7	100.0
	Writing and Foundations of Writing	6	7	100.0	6	9	100.0
7	Reading Prose and Poetry	7	8	100.0	7	10	100.0
	Reading Informational Text	8	9	100.0	8	11	100.0
	Vocabulary	4	5	100.0	4	7	100.0
	Writing and Foundations of Writing	6	7	100.0	6	9	100.0
8	Reading Prose and Poetry	7	8	100.0	7	10	100.0
	Reading Informational Text	8	9	100.0	8	11	100.0
	Vocabulary	4	5	100.0	4	7	100.0
	Writing and Foundations of Writing	6	7	100.0	6	9	100.0
Mathematics							
3	Number	10	10	90.8	11	12	99.4
	Algebra	5	5	99.7	6	7	98.0
	Geometry	7	7	86.9	8	9	99.7
	Data	5	5	96.6	6	7	96.2
4	Number	10	10	55.8	11	12	89.2
	Algebra	6	6	80.4	7	8	97.8
	Geometry	6	6	98.5	7	8	97.0
	Data	5	5	86.3	6	7	97.0
5	Number	10	10	76.3	11	12	94.7
	Algebra	6	6	96.3	7	8	96.9
	Geometry	6	6	83.1	7	8	98.1
	Data	5	5	94.8	6	7	90.4
6	Number	7	7	98.4	8	9	98.7
	Algebra	10	10	77.2	11	12	92.4
	Geometry	5	5	98.8	6	7	97.6
	Data	5	5	68.0	6	7	95.5
7	Number	6	6	93.0	7	8	89.4
	Algebra	9	9	78.3	10	11	94.1
	Geometry	5	5	81.1	6	7	89.3
	Data	7	7	96.6	8	9	98.8
8	Number	7	7	83.0	8	9	83.5
	Algebra	7	7	77.5	8	9	96.8
	Geometry	8	8	84.4	9	10	84.5
	Data	5	5	97.7	6	7	99.1

Table 5.5. Blueprint Constraint Accuracy by Reporting Category—Spring Simulations

Grade	Reporting Category	#Items			#Points		
		Min.	Max.	%Match	Min.	Max.	%Match
ELA							
3	Reading Prose and Poetry	7	9	100.0	7	11	100.0
	Reading Informational Text	7	9	100.0	7	11	100.0
	Vocabulary	4	5	100.0	4	7	100.0
	Writing and Foundations of Writing	6	7	100.0	6	9	100.0
4	Reading Prose and Poetry	7	9	100.0	7	11	100.0
	Reading Informational Text	7	9	100.0	7	11	100.0
	Vocabulary	4	5	100.0	4	7	100.0
	Writing and Foundations of Writing	6	7	100.0	6	9	100.0
5	Reading Prose and Poetry	7	9	100.0	7	11	100.0
	Reading Informational Text	7	9	100.0	7	11	100.0
	Vocabulary	4	5	100.0	4	7	100.0
	Writing and Foundations of Writing	6	7	100.0	6	9	100.0
6	Reading Prose and Poetry	7	9	100.0	7	11	100.0
	Reading Informational Text	7	9	100.0	7	11	100.0
	Vocabulary	4	5	100.0	4	7	100.0
	Writing and Foundations of Writing	6	7	100.0	6	9	100.0
7	Reading Prose and Poetry	7	8	100.0	7	10	100.0
	Reading Informational Text	8	9	100.0	8	11	100.0
	Vocabulary	4	5	100.0	4	7	100.0
	Writing and Foundations of Writing	6	7	100.0	6	9	100.0
8	Reading Prose and Poetry	7	8	100.0	7	10	100.0
	Reading Informational Text	8	9	100.0	8	11	100.0
	Vocabulary	4	5	100.0	4	7	100.0
	Writing and Foundations of Writing	6	7	100.0	6	9	100.0
Mathematics							
3	Number	10	10	100.0	11	12	100.0
	Algebra	5	5	100.0	6	7	100.0
	Geometry	7	7	100.0	8	9	100.0
	Data	5	5	100.0	6	7	100.0
4	Number	10	10	100.0	11	12	100.0
	Algebra	6	6	100.0	7	8	100.0
	Geometry	6	6	100.0	7	8	100.0
	Data	5	5	100.0	6	7	100.0
5	Number	10	10	100.0	11	12	100.0
	Algebra	6	6	100.0	7	8	100.0
	Geometry	6	6	100.0	7	8	100.0
	Data	5	5	100.0	6	7	100.0
6	Number	7	7	100.0	8	9	100.0
	Algebra	10	10	100.0	11	12	100.0
	Geometry	5	5	100.0	6	7	100.0
	Data	5	5	100.0	6	7	100.0
7	Number	6	6	100.0	7	8	100.0
	Algebra	9	9	100.0	10	11	100.0
	Geometry	5	5	100.0	6	7	100.0
	Data	7	7	100.0	8	9	100.0
8	Number	7	7	100.0	8	9	100.0
	Algebra	7	7	100.0	8	9	100.0
	Geometry	8	8	100.0	9	10	100.0
	Data	5	5	100.0	6	7	100.0

Table 5.6. Blueprint Constraint Accuracy by Reporting Category—Spring Engine Evaluation

Grade	Reporting Category	#Items			#Points		
		Min.	Max.	%Match	Min.	Max.	%Match
ELA							
3	Reading Prose and Poetry	7	9	100.0	7	11	100.0
	Reading Informational Text	7	9	100.0	7	11	100.0
	Vocabulary	4	5	100.0	4	7	100.0
	Writing and Foundations of Writing	6	7	100.0	6	9	100.0
4	Reading Prose and Poetry	7	9	100.0	7	11	100.0
	Reading Informational Text	7	9	100.0	7	11	100.0
	Vocabulary	4	5	100.0	4	7	100.0
	Writing and Foundations of Writing	6	7	100.0	6	9	100.0
5	Reading Prose and Poetry	7	9	100.0	7	11	100.0
	Reading Informational Text	7	9	100.0	7	11	100.0
	Vocabulary	4	5	100.0	4	7	100.0
	Writing and Foundations of Writing	6	7	100.0	6	9	100.0
6	Reading Prose and Poetry	7	9	100.0	7	11	100.0
	Reading Informational Text	7	9	100.0	7	11	100.0
	Vocabulary	4	5	100.0	4	7	100.0
	Writing and Foundations of Writing	6	7	100.0	6	9	100.0
7	Reading Prose and Poetry	7	8	100.0	7	10	100.0
	Reading Informational Text	8	9	100.0	8	11	100.0
	Vocabulary	4	5	100.0	4	7	100.0
	Writing and Foundations of Writing	6	7	100.0	6	9	100.0
8	Reading Prose and Poetry	7	8	100.0	7	10	100.0
	Reading Informational Text	8	9	100.0	8	11	100.0
	Vocabulary	4	5	100.0	4	7	100.0
	Writing and Foundations of Writing	6	7	100.0	6	9	100.0
Mathematics							
3	Number	10	10	100.0	11	12	100.0
	Algebra	5	5	100.0	6	7	100.0
	Geometry	7	7	100.0	8	9	100.0
	Data	5	5	100.0	6	7	100.0
4	Number	10	10	100.0	11	12	100.0
	Algebra	6	6	100.0	7	8	100.0
	Geometry	6	6	100.0	7	8	100.0
	Data	5	5	100.0	6	7	100.0
5	Number	10	10	100.0	11	12	100.0
	Algebra	6	6	100.0	7	8	100.0
	Geometry	6	6	100.0	7	8	100.0
	Data	5	5	100.0	6	7	93.2
6	Number	7	7	100.0	8	9	100.0
	Algebra	10	10	100.0	11	12	100.0
	Geometry	5	5	100.0	6	7	100.0
	Data	5	5	100.0	6	7	100.0
7	Number	6	6	100.0	7	8	97.9
	Algebra	9	9	100.0	10	11	100.0
	Geometry	5	5	100.0	6	7	100.0
	Data	7	7	100.0	8	9	100.0
8	Number	7	7	100.0	8	9	100.0
	Algebra	7	7	100.0	8	9	100.0
	Geometry	8	8	100.0	9	10	100.0
	Data	5	5	100.0	6	7	100.0

5.2.3. Item Exposure Rates

Table 5.7. Item Exposure Rates—Fall Simulations through Table 5.12 present the item exposure rates from the pre-administration simulation study and the post-administration engine evaluation study for fall, winter, and spring, respectively. Because different students receive different items based on blueprint constraints and their ability during an adaptive administration, it is ideal to have a low exposure rate. The exposure rate for each item is calculated as the percentage of students who received that item. For example, if Item 1 was administered to 500 out of 1,000 students, the exposure rate would be 50%. In the tables, **Error! Reference source not found.** “Total” is the total number of items in the operational item pool, and “Unused” shows the number and percentage of unused items that were never administered to students.

For the 2022–2023 administration, item exposure is being controlled by an update to a feature in the engine that assigns a weight to an item based on the number of times the item is seen by students. As the weight increases, that item is no longer preferred in the item-selection student-specific plan (SSP). This feature does not prevent the item from being seen by students if it is the best item in the pool to meet the requirements for that student; rather, this feature directs the engine to prefer additional items in the pool that might meet the requirements for the student over the item that has already been exposed. The results show that this updated feature, which has been applied since Spring 2021, combined with the new test design (i.e., including diagnostic items of adjacent grades), resulted in increased item pool usage (especially for ELA) compared with historical simulation results.

Table 5.7. Item Exposure Rates—Fall Simulations

Content Area	Grade	#Items				Exposure Rate											
						0–20%		21–40%		41–60%		61–80%		81–99%		100%	
		Total	Used	Unused	Unused %	N	%	N	%	N	%	N	%	N	%	N	%
ELA	3	660	519	141	21.36	490	94.41	26	5.01	3	0.58	0	0	0	0	0	0
	4	809	584	225	27.81	569	97.43	14	2.40	1	0.17	0	0	0	0	0	0
	5	787	584	203	25.79	569	97.43	14	2.40	1	0.17	0	0	0	0	0	0
	6	858	595	263	30.65	568	95.46	27	4.54	0	0	0	0	0	0	0	0
	7	1,126	466	660	58.61	428	91.85	32	6.87	6	1.29	0	0	0	0	0	0
	8	1,175	495	680	57.87	468	94.55	22	4.44	2	0.40	2	0.40	1	0.20	0	0
Math	3	2,247	868	1,379	61.37	851	98.04	12	1.38	5	0.58	0	0	0	0	0	0
	4	3,083	723	2,360	76.55	700	96.82	17	2.35	6	0.83	0	0	0	0	0	0
	5	3,777	676	3,101	82.10	658	97.34	13	1.92	5	0.74	0	0	0	0	0	0
	6	4,587	1,071	3,516	76.65	105	98.60	10	0.93	4	0.37	1	0.09	0	0	0	0
	7	5,449	910	4,539	83.30	894	98.24	11	1.21	5	0.55	0	0	0	0	0	0
	8	5,360	889	4,471	83.41	869	97.75	18	2.02	2	0.22	0	0	0	0	0	0

Table 5.8. Item Exposure Rates—Fall Engine Evaluation

Content Area	Grade	#Items				Exposure Rate											
						0–20%		21–40%		41–60%		61–80%		81–99%		100%	
		Total	Used	Unused	Unused %	N	%	N	%	N	%	N	%	N	%	N	%
ELA	3	660	520	140	21.21	493	94.81	24	4.62	3	0.58	0	0	0	0	0	0
	4	809	585	224	27.69	570	97.44	15	2.56	0	0	0	0	0	0	0	0
	5	787	510	277	35.20	502	98.43	6	1.18	2	0.39	0	0	0	0	0	0
	6	858	601	257	29.95	580	96.51	21	3.49	0	0	0	0	0	0	0	0
	7	1,126	473	653	57.99	440	93.02	29	6.13	4	0.85	0	0	0	0	0	0
	8	1,175	510	665	56.60	479	93.92	27	5.29	2	0.39	1	0.20	1	0.20	0	0
Math	3	2,247	808	1,439	64.04	790	97.77	13	1.61	5	0.62	0	0	0	0	0	0
	4	3,083	649	2,434	78.95	628	96.76	15	2.31	5	0.77	1	0.15	0	0	0	0
	5	3,777	520	3,257	86.23	502	96.54	13	2.50	5	0.96	0	0	0	0	0	0

Content Area	Grade	#Items				Exposure Rate											
						0–20%		21–40%		41–60%		61–80%		81–99%		100%	
		Total	Used	Unused	Unused %	N	%	N	%	N	%	N	%	N	%	N	%
	6	4,587	933	3,654	79.66	919	98.50	7	0.75	6	0.64	1	0.11	0	0	0	0
	7	5,449	749	4,700	86.25	737	98.40	5	0.67	6	0.80	1	0.13	0	0	0	0
	8	5,360	691	4,669	87.11	675	97.68	12	1.74	4	0.58	0	0	0	0	0	0

Table 5.9. Item Exposure Rates—Winter Simulations

Content Area	Grade	#Items				Exposure Rate											
						0–20%		21–40%		41–60%		61–80%		81–99%		100%	
		Total	Used	Unused	Unused %	N	%	N	%	N	%	N	%	N	%	N	%
ELA	3	548	422	126	22.99	393	93.13	17	4.03	11	2.61	1	0.24	0	0	0	0
	4	609	381	228	37.44	369	96.85	12	3.15	0	0	0	0	0	0	0	0
	5	614	385	229	37.30	354	91.95	29	7.53	2	0.52	0	0	0	0	0	0
	6	730	516	214	29.32	474	91.86	42	8.14	0	0	0	0	0	0	0	0
	7	940	380	560	59.57	360	94.74	7	1.84	13	3.42	0	0	0	0	0	0
	8	984	442	542	55.08	421	95.25	9	2.04	12	2.71	0	0	0	0	0	0
Math	3	2,925	871	2,054	70.22	857	98.39	14	1.61	0	0	0	0	0	0	0	0
	4	4,005	873	3,132	78.20	860	98.51	13	1.49	0	0	0	0	0	0	0	0
	5	4,832	706	4,126	85.39	690	97.73	16	2.27	0	0	0	0	0	0	0	0
	6	5,738	1,346	4,392	76.54	133	99.11	12	0.89	0	0	0	0	0	0	0	0
	7	6,814	1,231	5,583	81.93	121	98.94	13	1.06	0	0	0	0	0	0	0	0
	8	6,727	1,141	5,586	83.04	112	98.86	13	1.14	0	0	0	0	0	0	0	0

Table 5.10. Item Exposure Rates—Winter Engine Evaluation

Content Area	Grade	#Items				Exposure Rate											
						0–20%		21–40%		41–60%		61–80%		81–99%		100%	
		Total	Used	Unused	Unused %	N	%	N	%	N	%	N	%	N	%	N	%
ELA	3	548	422	126	22.99	393	93.13	16	3.79	12	2.84	1	0.24	0	0	0	0
	4	609	385	224	36.78	375	97.40	10	2.60	0	0	0	0	0	0	0	0

Content Area	Grade	#Items				Exposure Rate											
		Total	Used	Unused	Unused %	0–20%		21–40%		41–60%		61–80%		81–99%		100%	
						N	%	N	%	N	%	N	%	N	%	N	%
	5	614	393	221	35.99	362	92.11	30	7.63	1	0.25	0	0	0	0	0	0
	6	730	516	214	29.32	477	92.44	39	7.56	0	0	0	0	0	0	0	0
	7	940	380	560	59.57	362	95.26	5	1.32	13	3.42	0	0	0	0	0	0
	8	984	442	542	55.08	422	95.48	7	1.58	13	2.94	0	0	0	0	0	0
Math	3	2,925	935	1,990	68.03	921	98.50	12	1.28	2	0.21	0	0	0	0	0	0
	4	4,005	898	3,107	77.58	885	98.55	12	1.34	1	0.11	0	0	0	0	0	0
	5	4,832	702	4,130	85.47	687	97.86	15	2.14	0	0	0	0	0	0	0	0
	6	5,738	1,305	4,433	77.26	129	99.08	9	0.69	3	0.23	0	0	0	0	0	0
	7	6,814	1,232	5,582	81.92	121	98.94	11	0.89	2	0.16	0	0	0	0	0	0
	8	6,727	1,059	5,668	84.26	104	98.58	14	1.32	1	0.09	0	0	0	0	0	0

Table 5.11. Item Exposure Rates—Spring Simulations

Content Area	Grade	#Items				Exposure Rate											
		Total	Used	Unused	Unused %	0–20%		21–40%		41–60%		61–80%		81–99%		100%	
						N	%	N	%	N	%	N	%	N	%	N	%
ELA	3	924	653	271	29.33	653	100.00	0	0	0	0	0	0	0	0	0	0
	4	822	566	256	31.14	566	100.00	0	0	0	0	0	0	0	0	0	0
	5	882	598	284	32.20	596	99.67	2	0.33	0	0	0	0	0	0	0	0
	6	940	605	335	35.64	605	100.00	0	0	0	0	0	0	0	0	0	0
	7	1146	526	620	54.10	526	100.00	0	0	0	0	0	0	0	0	0	0
	8	1245	628	617	49.56	628	100.00	0	0	0	0	0	0	0	0	0	0
Math	3	3,236	1,425	1,811	55.96	1415	99.30	10	0.70	0	0	0	0	0	0	0	0
	4	4,218	966	3,252	77.10	957	99.07	9	0.93	0	0	0	0	0	0	0	0
	5	5,046	1,009	4,037	80.00	1004	99.50	5	0.50	0	0	0	0	0	0	0	0
	6	6,118	1,930	4,188	68.45	1924	99.69	5	0.26	1	0.05	0	0	0	0	0	0
	7	6,975	1,358	5,617	80.53	1350	99.41	8	0.59	0	0	0	0	0	0	0	0
	8	6,859	1,244	5,615	81.86	1241	99.76	2	0.16	1	0.08	0	0	0	0	0	0

Table 5.12. Item Exposure Rates—Spring Engine Evaluation

Content Area	Grade	#Items				Exposure Rate											
		Total	Used	Unused	Unused %	0–20%		21–40%		41–60%		61–80%		81–99%		100%	
						N	%	N	%	N	%	N	%	N	%	N	%
ELA	3	924	653	271	29.33	653	100.00	0	0	0	0	0	0	0	0	0	0
	4	822	566	256	31.14	566	100.00	0	0	0	0	0	0	0	0	0	0
	5	882	598	284	32.20	596	99.67	2	0.33	0	0	0	0	0	0	0	0
	6	940	605	335	35.64	605	100.00	0	0	0	0	0	0	0	0	0	0
	7	1,146	526	620	54.10	526	100.00	0	0	0	0	0	0	0	0	0	0
	8	1,245	628	617	49.56	628	100.00	0	0	0	0	0	0	0	0	0	0
Math	3	3,236	1,505	1,731	53.49	149	99.27	1	0.73	0	0	0	0	0	0	0	0
	4	4,218	1,138	3,080	73.02	112	98.77	1	1.23	0	0	0	0	0	0	0	0
	5	5,046	1,034	4,012	79.51	101	98.55	1	1.45	0	0	0	0	0	0	0	0
	6	6,118	2,052	4,066	66.46	203	99.22	1	0.78	0	0	0	0	0	0	0	0
	7	6,975	1,405	5,570	79.86	139	99.07	1	0.85	1	0.07	0	0	0	0	0	0
	8	6,859	1,307	5,552	80.94	129	98.93	1	0.99	1	0.08	0	0	0	0	0	0

5.2.4. Score Precision and Reliability

The analyses provided precision ability estimations that showed how well the adaptive engine recovered students' true ability based on the item pool. It included the standard deviation of estimated theta, mean SEM, SEM by deciles, and marginal reliability. The following indexes were used to examine the functionality of the engine during the simulations:

- Precision of ability estimation (how well the engine recovered students' true ability based on the item pool):
 - Bias: shows the difference between true and final estimated theta
 - *P* value for the z-test: determines if the difference of bias between the true and final estimated theta is statistically different. If the *p* value is larger than 0.05, there is no statistical difference of bias between the true and final estimated theta.
 - Root mean standard error (RMSE): provides the square of the bias statistic. While bias shows the difference between true and final estimated theta, RMSE shows the magnitude of the difference.
 - 95% and 99% coverage: shows the percentage of students who fall outside that range in terms of theta. Generally, it is expected that about 5% are outside the 95% confidence interval and about 1% are outside the 99% confidence interval.

Table 5.13 through Table 5.15 present the results of the precision ability estimation from the fall, winter, and spring simulations, respectively. Because this study did not involve an actual test administration, the adaptive engine is not scoring student responses but is instead simulating whether a student got items correct or incorrect based on the student's ability. Because a student's true theta is known, the engine should be able to recover the student's theta after administering all the items; this is the estimated theta. The null hypothesis is that there is no difference between true and estimated theta.

For the overall scores across all students, the mean biases are small (i.e., less than or equal to 0.03 in magnitude) for both ELA and mathematics, and the *p* value for the z-test supports the null hypothesis that there is no significant difference between the simulated students' true and final estimated thetas. For some reporting category scores across all students, the mean biases are larger, and the *p* value for the z-test results do not support the null hypothesis. This is because the number of items is much smaller at the reporting category level, and the large sample size could increase the likelihood of significant *p* values. The RMSE is also relatively small, showing that the engine typically recovered a value near the students' true theta.

Table 5.13. Mean Bias of the NSCAS Ability Estimation (True–Estimated)—Fall Simulations

Content Area	Grade	Reporting Category	Bias		<i>P</i> Value for Z-Test	RMSE	95% Coverage	99% Coverage
			Mean	SE				
ELA	3	Reading Vocabulary	-0.25	0.01	0.00	0.86	1.46	0.08
		Reading Comprehension	-0.16	0.01	0.00	0.59	4.93	0.99
		Writing Skills	0.21	0.01	0.00	0.81	1.39	0.19
		Overall	-0.08	0.00	0.00	0.44	6.35	1.29
	4	Reading Vocabulary	-0.23	0.01	0.00	0.93	0.88	0.01
		Reading Comprehension	-0.19	0.01	0.00	0.58	5.28	1.00
		Writing Skills	0.07	0.01	0.00	0.78	2.33	0.09
		Overall	-0.11	0.00	0.00	0.45	7.93	2.16
	5	Reading Vocabulary	-0.23	0.01	0.00	0.93	0.88	0.01
		Reading Comprehension	-0.19	0.01	0.00	0.58	5.28	1.00
		Writing Skills	0.07	0.01	0.00	0.78	2.33	0.09
		Overall	-0.11	0.00	0.00	0.45	7.93	2.16

Content Area	Grade	Reporting Category	Bias		P Value for Z-Test	RMSE	95% Coverage	99% Coverage
			Mean	SE				
	6	Overall	-0.11	0.00	0.00	0.45	7.93	2.16
		Reading Vocabulary	-0.07	0.01	0.00	0.89	0.91	0.03
		Reading Comprehension	0.00	0.01	0.98	0.53	3.10	0.28
		Writing Skills	-0.05	0.01	0.00	0.78	1.80	0.10
	7	Overall	-0.02	0.00	0.07	0.41	5.14	0.94
		Reading Vocabulary	-0.01	0.01	0.61	0.88	1.20	0.05
		Reading Comprehension	0.06	0.01	0.00	0.54	3.81	0.61
		Writing Skills	-0.02	0.01	0.19	0.78	1.68	0.04
	8	Overall	0.03	0.00	0.03	0.41	5.34	0.99
		Reading Vocabulary	-0.05	0.01	0.00	0.91	1.19	0.05
		Reading Comprehension	0.07	0.01	0.00	0.50	3.01	0.40
		Writing Skills	-0.09	0.01	0.00	0.79	1.96	0.06
Math	3	Overall	0.00	0.00	0.72	0.38	3.98	0.51
		Number	-0.15	0.01	0.00	0.62	4.46	0.73
		Algebra	-0.08	0.01	0.00	0.82	1.33	0.04
		Geometry	-0.16	0.01	0.00	0.75	2.25	0.21
		Data	-0.13	0.01	0.00	0.86	2.41	0.14
	4	Overall	-0.12	0.00	0.00	0.41	8.06	2.63
		Number	-0.05	0.01	0.00	0.63	3.10	0.34
		Algebra	0.00	0.01	0.85	0.80	1.43	0.05
		Geometry	-0.05	0.01	0.00	0.78	1.66	0.13
		Data	-0.01	0.01	0.27	0.89	1.46	0.09
	5	Overall	-0.03	0.00	0.00	0.40	6.53	1.54
		Number	0.01	0.01	0.26	0.61	3.06	0.31
		Algebra	0.03	0.01	0.01	0.80	1.71	0.09
		Geometry	0.01	0.01	0.48	0.79	1.06	0.04
		Data	0.01	0.01	0.61	0.86	0.69	0.01
	6	Overall	0.02	0.00	0.19	0.36	5.33	1.19
		Number	0.06	0.01	0.00	0.72	2.38	0.15
		Algebra	0.09	0.01	0.00	0.63	3.34	0.23
		Geometry	0.10	0.01	0.00	0.86	1.05	0.09
		Data	0.33	0.01	0.00	0.96	0.85	0.01
	7	Overall	0.12	0.00	0.00	0.39	7.00	1.68
		Number	0.22	0.01	0.00	0.84	1.56	0.04
		Algebra	0.16	0.01	0.00	0.67	3.15	0.30
		Geometry	0.27	0.01	0.00	0.89	1.09	0.01
		Data	0.22	0.01	0.00	0.78	2.31	0.20
	8	Overall	0.20	0.00	0.00	0.42	7.38	1.81
		Number	0.31	0.01	0.00	0.81	2.78	0.10
		Algebra	0.29	0.01	0.00	0.77	2.36	0.08
		Geometry	0.29	0.01	0.00	0.77	2.84	0.19
		Data	0.20	0.01	0.00	0.91	1.95	0.10
		Overall	0.27	0.00	0.00	0.45	8.78	2.05

Table 5.14. Mean Bias of the NSCAS Ability Estimation (True–Estimated)—Winter Simulations

Content Area	Grade	Reporting Category	Bias		P Value for Z-Test	RMSE	95% Coverage	99% Coverage
			Mean	SE				
ELA	3	Reading Prose and Poetry	-0.24	0.02	0.00	0.77	2.85	0.15
		Reading Informational Text	-0.25	0.02	0.00	0.83	4.55	0.65
		Vocabulary	-0.33	0.02	0.00	1.00	0.80	
		Writing and Foundations of Writing						
			0.16	0.02	0.00	0.96	2.30	0.10

Content Area	Grade	Reporting Category	Bias		P Value for Z-Test	RMSE	95% Coverage	99% Coverage
			Mean	SE				
		Overall	-0.17	0.01	0.00	0.46	8.30	1.90
	4	Reading Prose and Poetry	-0.20	0.02	0.00	0.83	1.92	0.05
		Reading Informational Text	-0.24	0.02	0.00	0.83	3.13	0.05
		Vocabulary	-0.25	0.02	0.00	1.03	0.66	
		Writing and Foundations of Writing	-0.01	0.02	0.84	0.91	2.32	0.10
		Overall	-0.15	0.01	0.00	0.47	7.86	1.76
	5	Reading Prose and Poetry	-0.18	0.02	0.00	0.85	1.16	0.05
		Reading Informational Text	-0.09	0.02	0.00	0.76	2.47	0.20
		Vocabulary	-0.17	0.02	0.00	1.00	0.45	
		Writing and Foundations of Writing	0.02	0.02	0.38	0.93	2.07	0.20
		Overall	-0.06	0.01	0.01	0.45	5.65	0.91
	6	Reading Prose and Poetry	-0.07	0.02	0.01	0.79	2.16	0.30
		Reading Informational Text	-0.02	0.02	0.43	0.73	1.81	0.10
		Vocabulary	-0.18	0.02	0.00	1.05	1.01	
		Writing and Foundations of Writing	-0.11	0.02	0.00	0.89	0.86	
		Overall	-0.05	0.01	0.07	0.43	5.64	0.60
	7	Reading Prose and Poetry	-0.01	0.02	0.71	0.81	1.46	0.05
		Reading Informational Text	-0.05	0.02	0.05	0.75	2.11	0.05
		Vocabulary	-0.06	0.02	0.02	0.98	0.45	0.05
		Writing and Foundations of Writing	-0.12	0.02	0.00	0.93	1.46	
		Overall	-0.03	0.01	0.25	0.43	3.82	0.50
	8	Reading Prose and Poetry	0.04	0.02	0.11	0.77	2.52	0.10
		Reading Informational Text	-0.08	0.02	0.00	0.70	1.81	0.05
		Vocabulary	-0.15	0.02	0.00	1.01	1.11	0.10
		Writing and Foundations of Writing	-0.13	0.02	0.00	0.93	1.26	0.05
		Overall	-0.04	0.01	0.10	0.40	4.39	0.71
Math	3	Number	-0.18	0.01	0.00	0.67	5.05	0.60
		Algebra	-0.19	0.02	0.00	0.93	1.90	0.20
		Geometry	-0.23	0.02	0.00	0.76	2.80	0.30
		Data	-0.16	0.02	0.00	0.89	3.15	0.45
		Overall	-0.19	0.01	0.00	0.45	11.3	3.45
	4	Number	-0.11	0.01	0.00	0.64	3.98	0.25
		Algebra	-0.05	0.02	0.02	0.83	2.67	0.15
		Geometry	-0.17	0.02	0.00	0.83	2.98	0.10
		Data	-0.09	0.02	0.00	0.92	2.22	
		Overall	-0.11	0.01	0.00	0.41	7.51	2.32
	5	Number	-0.05	0.01	0.04	0.63	3.13	0.15
		Algebra	-0.05	0.02	0.05	0.80	2.32	0.05
		Geometry	-0.03	0.02	0.26	0.84	1.82	0.10
		Data	-0.05	0.02	0.05	0.93	1.16	
		Overall	-0.05	0.01	0.05	0.40	6.81	1.56
	6	Number	0.07	0.02	0.00	0.76	3.02	0.50
		Algebra	0.05	0.01	0.02	0.67	4.08	0.35
		Geometry	0.08	0.02	0.00	0.89	1.26	0.20
		Data	0.20	0.02	0.00	0.92	0.86	0.05

Content Area	Grade	Reporting Category	Bias		P Value for Z-Test	RMSE	95% Coverage	99% Coverage
			Mean	SE				
	7	Overall	0.08	0.01	0.00	0.37	5.64	1.26
		Number	0.18	0.02	0.00	0.82	1.71	
		Algebra	0.16	0.02	0.00	0.71	3.78	0.35
		Geometry	0.24	0.02	0.00	0.94	1.56	
		Data	0.19	0.02	0.00	0.81	2.72	0.05
		Overall	0.18	0.01	0.00	0.41	6.96	1.31
	8	Number	0.31	0.02	0.00	0.82	3.63	0.15
		Algebra	0.25	0.02	0.00	0.79	2.22	0.25
		Geometry	0.25	0.02	0.00	0.76	3.23	0.15
		Data	0.22	0.02	0.00	0.93	1.86	0.10
		Overall	0.25	0.01	0.00	0.44	8.17	1.61

Table 5.15. Mean Bias of the NSCAS Ability Estimation (True–Estimated)—Spring Simulations

Content Area	Grade	Reporting Category	Bias		P Value for Z-Test	RMSE	95% Coverage	99% Coverage
			Mean	SE				
ELA	3	Reading Prose and Poetry	-0.09	0.02	0.00	0.75	2.10	0.10
		Reading Informational Text	-0.13	0.02	0.00	0.79	3.40	0.40
		Vocabulary	-0.19	0.02	0.00	0.99	0.55	0.05
		Writing and Foundations of Writing	0.08	0.02	0.01	0.87	1.30	0.00
		Overall	-0.07	0.01	0.02	0.44	6.10	1.35
	4	Reading Prose and Poetry	-0.10	0.02	0.00	0.77	1.70	0.05
		Reading Informational Text	-0.19	0.02	0.00	0.80	2.00	0.05
		Vocabulary	-0.21	0.02	0.00	1.00	0.45	0.00
		Writing and Foundations of Writing	-0.03	0.02	0.35	0.85	1.45	0.00
		Overall	-0.11	0.01	0.00	0.47	7.05	1.20
	5	Reading Prose and Poetry	-0.13	0.02	0.00	0.80	1.50	0.05
		Reading Informational Text	-0.08	0.02	0.00	0.79	2.45	0.10
		Vocabulary	-0.17	0.02	0.00	0.99	0.85	0.00
		Writing and Foundations of Writing	-0.07	0.02	0.01	0.86	2.00	0.00
		Overall	-0.07	0.01	0.01	0.44	6.20	1.55
	6	Reading Prose and Poetry	-0.09	0.02	0.00	0.80	1.55	0.05
		Reading Informational Text	-0.07	0.02	0.01	0.78	1.80	0.05
		Vocabulary	-0.15	0.02	0.00	0.97	0.55	0.15
		Writing and Foundations of Writing	-0.10	0.02	0.00	0.90	1.10	0.05
		Overall	-0.07	0.01	0.01	0.45	6.45	1.10
	7	Reading Prose and Poetry	-0.04	0.02	0.22	0.83	1.55	0.15
		Reading Informational Text	-0.06	0.02	0.05	0.75	1.50	0.05
		Vocabulary	-0.06	0.02	0.05	0.92	0.50	0.00
		Writing and Foundations of Writing	-0.14	0.02	0.00	0.85	0.90	0.05
		Overall	-0.04	0.01	0.19	0.43	5.00	1.00
	8	Reading Prose and Poetry	-0.09	0.02	0.01	0.80	1.55	0.05
		Reading Informational Text	-0.08	0.02	0.01	0.69	1.35	0.10

Content Area	Grade	Reporting Category	Bias		P Value for Z-Test	RMSE	95% Coverage	99% Coverage
			Mean	SE				
Math		Vocabulary	-0.14	0.02	0.00	0.96	0.55	0.05
		Writing and Foundations of Writing	-0.21	0.02	0.00	0.86	1.10	0.00
		Overall	-0.09	0.01	0.01	0.41	4.15	0.35
	3	Number	-0.25	0.01	0.00	0.66	5.65	1.00
		Algebra	-0.21	0.02	0.00	0.90	2.25	0.15
		Geometry	-0.26	0.02	0.00	0.78	3.20	0.30
		Data	-0.25	0.02	0.00	0.89	3.20	0.30
		Overall	-0.24	0.01	0.00	0.46	12.55	4.35
	4	Number	-0.17	0.01	0.00	0.68	5.30	0.80
		Algebra	-0.16	0.02	0.00	0.83	2.90	0.15
		Geometry	-0.20	0.02	0.00	0.82	3.40	0.15
		Data	-0.11	0.02	0.00	0.91	2.50	0.20
		Overall	-0.17	0.01	0.00	0.44	11.20	3.80
	5	Number	-0.05	0.01	0.09	0.65	4.20	0.45
		Algebra	-0.05	0.02	0.08	0.82	1.85	0.15
		Geometry	-0.02	0.02	0.55	0.79	1.95	0.20
		Data	-0.09	0.02	0.00	0.90	1.00	0.05
		Overall	-0.05	0.01	0.09	0.40	6.45	1.80
	6	Number	0.04	0.02	0.16	0.75	2.20	0.10
		Algebra	0.05	0.01	0.07	0.62	2.40	0.10
		Geometry	0.03	0.02	0.32	0.89	1.55	0.05
		Data	0.07	0.02	0.02	0.90	1.55	0.10
		Overall	0.04	0.01	0.17	0.37	6.30	1.85
	7	Number	0.08	0.02	0.01	0.81	1.20	0.20
		Algebra	0.16	0.01	0.00	0.67	2.70	0.15
		Geometry	0.17	0.02	0.00	0.89	1.55	0.00
		Data	0.14	0.02	0.00	0.77	2.35	0.00
		Overall	0.14	0.01	0.00	0.39	6.10	1.50
	8	Number	0.24	0.02	0.00	0.80	2.25	0.05
		Algebra	0.27	0.02	0.00	0.79	3.20	0.20
		Geometry	0.23	0.02	0.00	0.75	2.75	0.10
		Data	0.21	0.02	0.00	0.93	1.60	0.15
		Overall	0.23	0.01	0.00	0.45	8.85	2.70

Table 5.16 through Table 5.21 present the score precision and reliability estimates for the pre-administration simulation study and the post-administration engine evaluation study for fall, winter, and spring, respectively. Tables include the average number of items administered, the standard deviation (SD) of the estimated theta, the mean SEM, and a marginal reliability coefficient. The SD and mean SEM are relatively small, and the range of the marginal reliability for the overall scores is close to or higher than 0.90. These results indicate that, overall, the score precision is reasonable: The overall mean SEM values are approximately 0.40, while the reliability estimates are consistent with the guidelines for reliability in a graduation test (Phillips & Camara, 2006). The reliability for the overall scores shows higher reliability estimates compared with that of reporting category scores, which can be expected as more items are contributing to the overall scores.

Table 5.16. Score Precision & Reliability, Items Contributing to NSCAS—Fall Simulations

Content Area	Grade	Reporting Category	Mean #Items	SD of Estimated Theta	Mean SEM	Reliability
ELA	3	Reading Vocabulary	6.0	1.58	0.90	0.66
		Reading Comprehension	14.0	1.46	0.59	0.83
		Writing Skills	7.0	1.51	0.86	0.66
		Overall	27.0	1.42	0.41	0.91
	4	Reading Vocabulary	6.0	1.61	0.95	0.63
		Reading Comprehension	14.0	1.42	0.54	0.85
		Writing Skills	7.0	1.53	0.76	0.75
		Overall	27.0	1.39	0.38	0.92
	5	Reading Vocabulary	6.0	1.61	0.95	0.63
		Reading Comprehension	14.0	1.42	0.54	0.85
		Writing Skills	7.0	1.53	0.76	0.75
		Overall	27.0	1.39	0.38	0.92
	6	Reading Vocabulary	6.0	1.54	0.94	0.60
		Reading Comprehension	14.0	1.36	0.53	0.84
		Writing Skills	7.0	1.56	0.73	0.77
		Overall	27.0	1.34	0.37	0.92
	7	Reading Vocabulary	6.0	1.52	0.90	0.62
		Reading Comprehension	14.0	1.36	0.53	0.85
		Writing Skills	7.0	1.56	0.76	0.75
		Overall	27.0	1.33	0.37	0.92
	8	Reading Vocabulary	6.0	1.55	0.97	0.57
		Reading Comprehension	14.0	1.31	0.52	0.84
		Writing Skills	7.0	1.54	0.76	0.75
		Overall	27.0	1.30	0.38	0.92
Math	3	Number	10.0	1.18	0.62	0.72
		Algebra	5.0	1.29	0.86	0.54
		Geometry	7.0	1.25	0.74	0.64
		Data	5.0	1.29	0.93	0.47
		Overall	26.0	1.10	0.36	0.89
	4	Number	9.0	1.26	0.64	0.73
		Algebra	6.0	1.36	0.81	0.64
		Geometry	6.0	1.36	0.78	0.65
		Data	5.0	1.43	0.92	0.57
		Overall	26.0	1.19	0.36	0.91
	5	Number	10.0	1.33	0.63	0.77
		Algebra	6.0	1.41	0.80	0.67
		Geometry	6.0	1.37	0.78	0.65
		Data	5.0	1.46	0.86	0.64
		Overall	26.0	1.22	0.35	0.92
	6	Number	7.0	1.38	0.72	0.72
		Algebra	10.0	1.32	0.64	0.76
		Geometry	5.0	1.43	0.86	0.62
		Data	4.0	1.53	0.92	0.62
		Overall	26.0	1.24	0.36	0.92
	7	Number	6.0	1.51	0.84	0.67
		Algebra	9.0	1.42	0.67	0.77
		Geometry	5.0	1.55	0.86	0.67
		Data	7.0	1.44	0.76	0.71
		Overall	26.0	1.32	0.36	0.92
	8	Number	7.0	1.53	0.78	0.73
		Algebra	7.0	1.55	0.74	0.76
		Geometry	8.0	1.50	0.72	0.76
		Data	5.0	1.59	0.90	0.67

Content Area	Grade	Reporting Category	Mean #Items	SD of Estimated Theta	Mean SEM	Reliability
		Overall	26.0	1.39	0.36	0.93

Table 5.17. Score Precision & Reliability, Items Contributing to NSCAS—Fall Engine Evaluation

Content Area	Grade	Reporting Category	Mean #Items	SD of Estimated Theta	Mean SEM	Reliability
ELA	3	Reading Vocabulary	6.0	1.49	0.89	0.62
		Reading Comprehension	14.0	1.28	0.56	0.80
		Writing Skills	7.0	1.34	0.80	0.64
		Overall	27.0	1.19	0.39	0.89
	4	Reading Vocabulary	6.0	1.50	0.93	0.59
		Reading Comprehension	14.0	1.31	0.52	0.84
		Writing Skills	7.0	1.28	0.75	0.65
		Overall	27.0	1.21	0.37	0.91
	5	Reading Vocabulary	6.0	1.55	0.95	0.60
		Reading Comprehension	14.0	1.27	0.52	0.83
		Writing Skills	7.0	1.33	0.74	0.68
		Overall	27.0	1.20	0.37	0.91
	6	Reading Vocabulary	6.0	1.38	0.93	0.53
		Reading Comprehension	14.0	1.17	0.51	0.81
		Writing Skills	7.0	1.28	0.73	0.65
		Overall	27.0	1.10	0.36	0.89
	7	Reading Vocabulary	6.0	1.36	0.85	0.58
		Reading Comprehension	14.0	1.24	0.51	0.83
		Writing Skills	7.0	1.15	0.74	0.57
		Overall	27.0	1.07	0.36	0.89
	8	Reading Vocabulary	6.0	1.33	0.93	0.49
		Reading Comprehension	14.0	1.10	0.51	0.79
		Writing Skills	7.0	1.11	0.76	0.51
		Overall	27.0	0.96	0.36	0.86
Math	3	Number	9.9	1.17	0.62	0.71
		Algebra	5.0	1.43	0.89	0.60
		Geometry	6.6	1.27	0.75	0.63
		Data	4.9	1.46	0.94	0.57
		Overall	26.0	1.13	0.36	0.90
	4	Number	9.5	1.15	0.65	0.67
		Algebra	5.7	1.44	0.82	0.67
		Geometry	6.0	1.42	0.83	0.64
		Data	4.9	1.60	0.95	0.63
		Overall	26.0	1.11	0.36	0.90
	5	Number	9.8	1.26	0.63	0.74
		Algebra	5.9	1.35	0.84	0.59
		Geometry	5.9	1.42	0.80	0.66
		Data	5.0	1.55	0.91	0.63
		Overall	27.0	1.12	0.35	0.90
	6	Number	7.0	1.34	0.74	0.69
		Algebra	9.6	1.28	0.65	0.74
		Geometry	5.0	1.39	0.90	0.56
		Data	4.4	1.40	0.99	0.46
		Overall	26.0	1.12	0.36	0.89
	7	Number	5.9	1.44	0.83	0.65
		Algebra	8.6	1.28	0.67	0.72
		Geometry	4.9	1.34	0.94	0.47
		Data	7.0	1.35	0.78	0.65

Content Area	Grade	Reporting Category	Mean #Items	SD of Estimated Theta	Mean SEM	Reliability
	8	Overall	26.0	1.11	0.36	0.89
		Number	6.9	1.43	0.80	0.67
		Algebra	6.7	1.46	0.77	0.71
		Geometry	7.7	1.35	0.76	0.67
		Data	5.0	1.43	0.91	0.58
		Overall	26.0	1.19	0.37	0.90

Table 5.18. Score Precision & Reliability, Items Contributing to NSCAS—Winter Simulations

Content Area	Grade	Reporting Category	#Items		SD of Estimated Theta	Mean SEM	Reliability
			Mean	SD			
ELA	3	Reading Prose and Poetry	8.7	0.7	1.58	0.73	0.78
		Reading Informational Text	9.0	0.1	1.61	0.80	0.74
		Vocabulary	5.0	0.1	1.70	1.01	0.61
		Writing and Foundations of Writing	6.0	0.0	1.63	1.09	0.48
		Overall	29.0	0.7	1.44	0.40	0.92
	4	Reading Prose and Poetry	7.7	0.8	1.55	0.80	0.72
		Reading Informational Text	8.6	0.7	1.53	0.76	0.74
		Vocabulary	5.0	0.1	1.62	1.13	0.47
		Writing and Foundations of Writing	6.0	0.2	1.57	0.86	0.69
		Overall	27.0	0.8	1.36	0.40	0.91
	5	Reading Prose and Poetry	7.6	0.8	1.48	0.84	0.66
		Reading Informational Text	8.8	0.5	1.43	0.75	0.72
		Vocabulary	4.9	0.2	1.56	1.10	0.46
		Writing and Foundations of Writing	6.0	0.0	1.56	0.95	0.60
		Overall	27.0	0.6	1.30	0.41	0.90
	6	Reading Prose and Poetry	8.3	0.9	1.49	0.79	0.70
		Reading Informational Text	9.0	0.2	1.46	0.71	0.75
		Vocabulary	5.0	0.0	1.63	1.12	0.49
		Writing and Foundations of Writing	6.0	0.0	1.59	0.87	0.68
		Overall	28.0	0.9	1.33	0.38	0.91
	7	Reading Prose and Poetry	7.7	0.5	1.47	0.82	0.66
		Reading Informational Text	8.9	0.2	1.48	0.74	0.73
		Vocabulary	5.0	0.1	1.56	1.05	0.49
		Writing and Foundations of Writing	6.0	0.0	1.58	0.97	0.57
		Overall	28.0	0.5	1.33	0.40	0.90
	8	Reading Prose and Poetry	7.8	0.4	1.44	0.80	0.68
		Reading Informational Text	9.0	0.1	1.40	0.70	0.74
		Vocabulary	5.0	0.2	1.54	1.12	0.40
		Writing and Foundations of Writing	6.0	0.0	1.60	0.96	0.61
		Overall	28.0	0.5	1.30	0.39	0.91
Math	3	Number	9.9	0.3	1.31	0.62	0.77
		Algebra	5.0	0.1	1.44	0.87	0.62
		Geometry	6.8	0.4	1.34	0.74	0.69
		Data	5.0	0.2	1.40	0.88	0.58
		Overall	27.0	0.6	1.20	0.35	0.91
	4	Number	9.6	0.9	1.30	0.62	0.76
		Algebra	5.9	0.4	1.44	0.80	0.68
		Geometry	6.0	0.2	1.41	0.80	0.66
		Data	4.9	0.3	1.48	0.91	0.60
		Overall	26.0	1.6	1.22	0.36	0.91
	5	Number	9.6	0.8	1.42	0.63	0.80
		Algebra	6.0	0.3	1.48	0.80	0.70

Content Area	Grade	Reporting Category	#Items		SD of Estimated Theta	Mean SEM	Reliability
			Mean	SD			
		Geometry	5.8	0.5	1.50	0.80	0.70
		Data	4.9	0.3	1.58	0.90	0.64
		Overall	26.0	1.4	1.32	0.35	0.93
	6	Number	7.0	0.2	1.41	0.73	0.72
		Algebra	9.7	0.7	1.34	0.65	0.76
		Geometry	5.0	0.2	1.45	0.88	0.61
		Data	4.8	0.4	1.47	0.90	0.60
		Overall	26.0	1.3	1.22	0.36	0.91
	7	Number	5.9	0.5	1.52	0.84	0.68
		Algebra	8.7	0.9	1.47	0.68	0.78
		Geometry	4.9	0.4	1.64	0.90	0.67
		Data	7.0	0.3	1.48	0.76	0.73
		Overall	26.0	1.9	1.35	0.36	0.93
	8	Number	6.8	0.7	1.58	0.79	0.74
		Algebra	6.8	0.7	1.63	0.75	0.78
		Geometry	7.7	0.9	1.57	0.73	0.77
		Data	4.9	0.3	1.67	0.90	0.70
		Overall	26.0	2.4	1.46	0.37	0.93

Table 5.19. Score Precision & Reliability, Items Contributing to NSCAS—Winter Engine Evaluation

Content Area	Grade	Reporting Category	#Items		SD of Estimated Theta	Mean SEM	Reliability
			Mean	SD			
ELA	3	Reading Prose and Poetry	8.7	0.7	1.51	0.73	0.75
		Reading Informational Text	9.0	0.1	1.36	0.76	0.68
		Vocabulary	6.0	0.0	1.57	1.02	0.55
		Writing and Foundations of Writing	5.0	0.1	1.38	0.93	0.52
		Overall	29.0	0.7	1.19	0.39	0.89
	4	Reading Prose and Poetry	7.7	0.8	1.45	0.79	0.69
		Reading Informational Text	8.7	0.7	1.49	0.74	0.74
		Vocabulary	6.0	0.2	1.52	1.12	0.41
		Writing and Foundations of Writing	5.0	0.1	1.41	0.85	0.62
		Overall	27.0	0.7	1.21	0.39	0.90
	5	Reading Prose and Poetry	7.6	0.8	1.37	0.82	0.62
		Reading Informational Text	8.8	0.5	1.34	0.73	0.69
		Vocabulary	6.0	0.0	1.47	1.06	0.44
		Writing and Foundations of Writing	5.0	0.2	1.45	0.92	0.58
		Overall	27.0	0.6	1.13	0.40	0.88
	6	Reading Prose and Poetry	8.5	0.9	1.42	0.77	0.69
		Reading Informational Text	9.0	0.2	1.27	0.70	0.69
		Vocabulary	6.0	0.0	1.46	1.09	0.41
		Writing and Foundations of Writing	5.0	0.1	1.25	0.80	0.56
		Overall	28.0	0.9	1.07	0.37	0.88
	7	Reading Prose and Poetry	7.7	0.5	1.40	0.79	0.66
		Reading Informational Text	9.0	0.1	1.32	0.71	0.70
		Vocabulary	6.0	0.0	1.45	1.00	0.48
		Writing and Foundations of Writing	5.0	0.1	1.16	0.86	0.41
		Overall	28.0	0.5	1.06	0.38	0.87
	8	Reading Prose and Poetry	7.8	0.4	1.42	0.79	0.68
		Reading Informational Text	9.0	0.1	1.24	0.68	0.69
		Vocabulary	6.0	0.0	1.41	1.05	0.39
		Writing and Foundations of Writing	5.0	0.1	1.17	0.90	0.37

Content Area	Grade	Reporting Category	#Items		SD of Estimated Theta	Mean SEM	Reliability
			Mean	SD			
		Overall	28.0	0.4	1.03	0.38	0.87
Math	3	Number	9.9	0.3	1.33	0.62	0.78
		Algebra	5.0	0.1	1.69	0.91	0.69
		Geometry	6.9	0.3	1.36	0.74	0.70
		Data	5.0	0.2	1.64	0.93	0.66
		Overall	27.0	0.6	1.27	0.35	0.92
	4	Number	9.4	0.9	1.37	0.64	0.77
		Algebra	5.8	0.5	1.61	0.82	0.73
		Geometry	6.0	0.1	1.62	0.84	0.72
		Data	4.8	0.4	1.67	0.95	0.66
		Overall	26.0	1.6	1.29	0.35	0.92
	5	Number	9.7	0.8	1.47	0.63	0.81
		Algebra	6.0	0.2	1.47	0.84	0.66
		Geometry	5.8	0.4	1.61	0.82	0.72
		Data	4.9	0.2	1.72	0.94	0.68
		Overall	26.0	1.1	1.32	0.35	0.93
	6	Number	7.0	0.2	1.53	0.75	0.75
		Algebra	9.7	0.8	1.44	0.65	0.79
		Geometry	5.0	0.1	1.49	0.90	0.60
		Data	4.7	0.5	1.43	1.00	0.47
		Overall	26.0	1.4	1.25	0.36	0.92
	7	Number	5.9	0.4	1.56	0.85	0.69
		Algebra	8.7	0.8	1.43	0.68	0.77
		Geometry	4.8	0.4	1.55	0.97	0.57
		Data	7.0	0.2	1.43	0.79	0.68
		Overall	26.0	1.5	1.25	0.36	0.92
	8	Number	6.7	0.7	1.52	0.83	0.69
		Algebra	6.7	0.6	1.60	0.77	0.76
		Geometry	7.7	0.8	1.52	0.76	0.74
		Data	5.0	0.2	1.54	0.92	0.63
		Overall	26.0	2.0	1.32	0.37	0.92

Table 5.20. Score Precision & Reliability, Items Contributing to NSCAS—Spring Simulations

Content Area	Grade	Reporting Category	#Items		SD of Estimated Theta	Mean SEM	Reliability
			Mean	SD			
ELA	3	Reading Prose and Poetry	7.7	0.8	1.57	0.79	0.73
		Reading Informational Text	8.7	0.6	1.65	0.77	0.77
		Vocabulary	4.9	0.3	1.74	1.05	0.61
		Writing and Foundations of Writing	6.0	0.1	1.63	0.93	0.66
		Overall	27.3	0.6	1.49	0.40	0.92
	4	Reading Prose and Poetry	8.1	0.8	1.63	0.76	0.77
		Reading Informational Text	8.0	0.8	1.66	0.77	0.77
		Vocabulary	5.0	0.2	1.73	1.11	0.55
		Writing and Foundations of Writing	6.0	0.1	1.66	0.87	0.72
		Overall	27.0	0.2	1.51	0.40	0.93
	5	Reading Prose and Poetry	7.8	0.9	1.55	0.79	0.72
		Reading Informational Text	8.5	0.8	1.54	0.73	0.76
		Vocabulary	4.9	0.3	1.64	1.06	0.54
		Writing and Foundations of Writing	6.0	0.0	1.60	0.87	0.69
		Overall	27.2	0.4	1.39	0.39	0.92
	6	Reading Prose and Poetry	7.6	0.7	1.63	0.79	0.74

Content Area	Grade	Reporting Category	#Items		SD of Estimated Theta	Mean SEM	Reliability
			Mean	SD			
		Reading Informational Text	8.6	0.6	1.63	0.74	0.78
		Vocabulary	4.9	0.2	1.69	1.06	0.57
		Writing and Foundations of Writing	6.0	0.1	1.73	0.86	0.74
		Overall	27.1	0.4	1.49	0.38	0.93
	7	Reading Prose and Poetry	7.1	0.3	1.68	0.85	0.72
		Reading Informational Text	8.9	0.2	1.65	0.73	0.79
		Vocabulary	5.0	0.1	1.69	1.03	0.59
		Writing and Foundations of Writing	6.0	0.1	1.68	0.85	0.73
		Overall	27.0	0.1	1.50	0.39	0.93
	8	Reading Prose and Poetry	7.3	0.5	1.66	0.81	0.74
		Reading Informational Text	8.8	0.4	1.61	0.72	0.79
		Vocabulary	5.0	0.2	1.74	1.09	0.56
		Writing and Foundations of Writing	6.0	0.1	1.71	0.85	0.73
		Overall	27.0	0.2	1.51	0.39	0.93
Math	3	Number	10	0	1.39	0.62	0.80
		Algebra	5	0	1.52	0.87	0.67
		Geometry	7	0	1.45	0.73	0.74
		Data	5	0	1.54	0.87	0.67
		Overall	27	0	1.31	0.35	0.93
	4	Number	10	0	1.52	0.63	0.83
		Algebra	6	0	1.60	0.80	0.74
		Geometry	6	0	1.59	0.79	0.75
		Data	5	0	1.66	0.90	0.69
		Overall	27	0	1.44	0.36	0.94
	5	Number	10	0	1.67	0.63	0.86
		Algebra	6	0	1.74	0.81	0.78
		Geometry	6	0	1.68	0.79	0.77
		Data	5	0	1.79	0.91	0.72
		Overall	27	0	1.57	0.36	0.95
	6	Number	7	0	1.59	0.72	0.79
		Algebra	10	0	1.49	0.62	0.82
		Geometry	5	0	1.63	0.90	0.68
		Data	5	0	1.69	0.87	0.73
		Overall	27	0	1.43	0.35	0.94
	7	Number	6	0	1.70	0.83	0.75
		Algebra	9	0	1.66	0.66	0.84
		Geometry	5	0	1.77	0.88	0.74
		Data	7	0	1.64	0.75	0.79
		Overall	27	0	1.54	0.36	0.95
	8	Number	7	0	1.76	0.78	0.80
		Algebra	7	0	1.78	0.74	0.82
		Geometry	8	0	1.77	0.73	0.83
		Data	5	0	1.85	0.91	0.75
		Overall	27	0	1.66	0.37	0.95

Table 5.21. Score Precision & Reliability, Items Contributing to NSCAS—Spring Engine Evaluation

Content Area	Grade	Reporting Category	#Items		SD of Estimated Theta	Mean SEM	Reliability
			Mean	SD			
ELA	3	Reading Prose and Poetry	7.8	0.8	1.47	0.76	0.72
		Reading Informational Text	8.7	0.6	1.47	0.72	0.75
		Vocabulary	6.0	0.1	1.61	1.06	0.53

Content Area	Grade	Reporting Category	#Items		SD of Estimated Theta	Mean SEM	Reliability
			Mean	SD			
		Writing and Foundations of Writing Overall	4.9 27.0	0.2 0.8	1.47 1.25	0.86 0.38	0.65 0.91
	4	Reading Prose and Poetry	8.1	0.8	1.43	0.74	0.71
		Reading Informational Text	7.9	0.8	1.50	0.76	0.73
		Vocabulary	6.0	0.1	1.59	1.15	0.43
		Writing and Foundations of Writing	5.0	0.1	1.54	0.86	0.67
		Overall	27.0	0.4	1.27	0.38	0.91
	5	Reading Prose and Poetry	7.9	0.9	1.44	0.77	0.69
		Reading Informational Text	8.5	0.8	1.40	0.72	0.73
		Vocabulary	6.0	0.2	1.54	1.05	0.49
		Writing and Foundations of Writing	4.9	0.3	1.47	0.85	0.65
		Overall	27.0	0.8	1.21	0.38	0.90
	6	Reading Prose and Poetry	7.4	0.7	1.38	0.75	0.69
		Reading Informational Text	8.7	0.6	1.36	0.70	0.73
		Vocabulary	6.0	0.1	1.47	1.05	0.45
		Writing and Foundations of Writing	5.0	0.2	1.35	0.81	0.62
		Overall	27.0	0.7	1.13	0.36	0.90
	7	Reading Prose and Poetry	7.0	0.2	1.41	0.80	0.66
		Reading Informational Text	9.0	0.2	1.34	0.69	0.72
		Vocabulary	6.0	0.2	1.47	1.00	0.49
		Writing and Foundations of Writing	5.0	0.1	1.35	0.81	0.62
		Overall	27.0	0.6	1.13	0.37	0.89
	8	Reading Prose and Poetry	7.4	0.5	1.40	0.76	0.68
		Reading Informational Text	8.7	0.5	1.28	0.68	0.71
		Vocabulary	6.0	0.1	1.46	1.06	0.41
		Writing and Foundations of Writing	5.0	0.2	1.34	0.81	0.62
		Overall	27.0	0.7	1.11	0.36	0.89
Math	3	Number	10.0	0.1	1.73	0.65	0.86
		Algebra	5.0	0.1	1.83	0.90	0.75
		Geometry	7.0	0.1	1.71	0.75	0.80
		Data	5.0	0.1	1.91	0.90	0.77
		Overall	27.0	0.3	1.61	0.36	0.95
	4	Number	10.0	0.1	1.67	0.64	0.85
		Algebra	6.0	0.1	1.83	0.80	0.80
		Geometry	6.0	0.0	1.85	0.81	0.80
		Data	5.0	0.1	1.92	0.92	0.76
		Overall	27.0	0.3	1.59	0.35	0.95
	5	Number	10.0	0.2	1.66	0.63	0.85
		Algebra	6.0	0.1	1.78	0.81	0.79
		Geometry	6.0	0.1	1.72	0.79	0.78
		Data	5.0	0.1	1.79	0.92	0.72
		Overall	27.0	0.4	1.52	0.35	0.95
	6	Number	7.0	0.1	1.71	0.73	0.81
		Algebra	10.0	0.2	1.66	0.62	0.86
		Geometry	5.0	0.1	1.76	0.90	0.72
		Data	5.0	0.1	1.85	0.88	0.76
		Overall	27.0	0.6	1.56	0.35	0.95
	7	Number	6.0	0.1	1.71	0.82	0.76
		Algebra	9.0	0.2	1.63	0.66	0.84
		Geometry	5.0	0.1	1.86	0.89	0.76
		Data	7.0	0.1	1.68	0.76	0.79
		Overall	27.0	0.5	1.52	0.35	0.95
	8	Number	7.0	0.1	1.73	0.78	0.79

Content Area	Grade	Reporting Category	#Items		SD of Estimated Theta	Mean SEM	Reliability
			Mean	SD			
		Algebra	7.0	0.1	1.80	0.74	0.82
		Geometry	8.0	0.2	1.83	0.72	0.84
		Data	5.0	0.1	1.77	0.90	0.73
		Overall	27.0	0.6	1.61	0.36	0.95

Table 5.22 through Table 5.27 present the average SEM by decile of the overall proficiency score, including the overall student ability distribution, for the fall pre-administration simulation study and the post-administration engine evaluation study for fall, winter, and spring, respectively. A decile is similar to a percentile rank, with 10 ranks related to the 10th, 20th, . . . 90th, and 100th percentile ranks. The average SEM is similar across deciles, except Decile 1 and Decile 10, which have a higher standard error compared with the other deciles. Overall, the SEM is within acceptable ranges (i.e., less than 0.40).

Table 5.22. SEM by Deciles for NSCAS Scores—Fall Simulations

Content Area	Grade	Proficiency Score Distribution										Overall
		Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10	
ELA	3	0.55	0.50	0.44	0.41	0.38	0.37	0.36	0.36	0.36	0.38	0.41
	4	0.46	0.41	0.39	0.38	0.37	0.36	0.35	0.35	0.36	0.40	0.38
	5	0.46	0.41	0.39	0.38	0.37	0.36	0.35	0.35	0.36	0.40	0.38
	6	0.48	0.39	0.36	0.35	0.35	0.34	0.35	0.35	0.36	0.40	0.37
	7	0.47	0.37	0.35	0.35	0.35	0.35	0.35	0.35	0.36	0.42	0.37
	8	0.45	0.38	0.36	0.35	0.35	0.35	0.35	0.36	0.37	0.43	0.38
Math	3	0.39	0.37	0.37	0.36	0.36	0.35	0.35	0.35	0.34	0.35	0.36
	4	0.48	0.39	0.36	0.35	0.34	0.34	0.34	0.34	0.34	0.35	0.36
	5	0.41	0.36	0.34	0.34	0.34	0.34	0.34	0.34	0.34	0.35	0.35
	6	0.41	0.37	0.36	0.35	0.35	0.35	0.34	0.35	0.35	0.35	0.36
	7	0.46	0.38	0.36	0.35	0.34	0.34	0.34	0.34	0.34	0.35	0.36
	8	0.47	0.38	0.37	0.36	0.35	0.35	0.34	0.34	0.34	0.35	0.36

Table 5.23. SEM by Deciles for NSCAS Scores—Fall Engine Evaluation

Content Area	Grade	Proficiency Score Distribution										Overall
		Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10	
ELA	3	0.48	0.42	0.40	0.38	0.37	0.36	0.36	0.36	0.36	0.38	0.39
	4	0.42	0.39	0.38	0.37	0.35	0.35	0.35	0.35	0.35	0.38	0.37
	5	0.42	0.39	0.37	0.36	0.36	0.35	0.35	0.34	0.35	0.38	0.37
	6	0.44	0.39	0.37	0.35	0.34	0.34	0.34	0.34	0.35	0.35	0.36
	7	0.43	0.37	0.35	0.35	0.35	0.34	0.34	0.33	0.33	0.36	0.36
	8	0.42	0.38	0.36	0.35	0.35	0.35	0.35	0.35	0.35	0.36	0.36
Math	3	0.38	0.37	0.37	0.36	0.35	0.35	0.35	0.34	0.34	0.36	0.36
	4	0.43	0.38	0.36	0.35	0.34	0.34	0.34	0.34	0.34	0.35	0.36
	5	0.39	0.36	0.34	0.34	0.34	0.34	0.34	0.34	0.33	0.35	0.35
	6	0.41	0.38	0.37	0.36	0.36	0.35	0.35	0.35	0.35	0.35	0.36

Content Area	Grade	Proficiency Score Distribution										Overall
		Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10	
	7	0.42	0.38	0.37	0.36	0.35	0.35	0.34	0.34	0.34	0.35	0.36
	8	0.45	0.39	0.38	0.37	0.37	0.36	0.35	0.34	0.34	0.34	0.37

Table 5.24. SEM by Deciles for NSCAS Scores—Winter Simulations

Content Area	Grade	Proficiency Score Distribution										Overall
		Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10	
ELA	3	0.55	0.45	0.40	0.37	0.37	0.36	0.36	0.36	0.38	0.42	0.40
	4	0.47	0.40	0.38	0.37	0.37	0.37	0.37	0.38	0.40	0.46	0.40
	5	0.45	0.40	0.39	0.39	0.39	0.38	0.39	0.39	0.41	0.48	0.41
	6	0.45	0.38	0.36	0.35	0.35	0.34	0.36	0.37	0.40	0.47	0.38
	7	0.46	0.39	0.36	0.36	0.35	0.35	0.36	0.37	0.41	0.59	0.40
	8	0.43	0.37	0.36	0.36	0.36	0.36	0.37	0.38	0.41	0.50	0.39
Math	3	0.39	0.37	0.36	0.35	0.34	0.34	0.34	0.34	0.34	0.36	0.35
	4	0.44	0.36	0.35	0.34	0.34	0.34	0.34	0.35	0.35	0.36	0.36
	5	0.40	0.35	0.35	0.34	0.34	0.34	0.34	0.34	0.35	0.38	0.35
	6	0.41	0.37	0.36	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.36
	7	0.46	0.38	0.36	0.35	0.35	0.34	0.35	0.35	0.35	0.36	0.36
	8	0.49	0.38	0.37	0.36	0.35	0.35	0.34	0.34	0.34	0.36	0.37

Table 5.25. SEM by Deciles for NSCAS Scores—Winter Engine Evaluation

Content Area	Grade	Proficiency Score Distribution										Overall
		Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10	
ELA	3	0.47	0.41	0.38	0.36	0.36	0.36	0.36	0.37	0.39	0.42	0.39
	4	0.43	0.39	0.38	0.37	0.37	0.36	0.37	0.37	0.39	0.46	0.39
	5	0.42	0.40	0.39	0.39	0.38	0.38	0.38	0.38	0.39	0.44	0.40
	6	0.42	0.39	0.37	0.36	0.35	0.34	0.34	0.35	0.36	0.41	0.37
	7	0.45	0.39	0.37	0.37	0.36	0.36	0.35	0.35	0.36	0.42	0.38
	8	0.42	0.38	0.37	0.36	0.36	0.36	0.36	0.36	0.37	0.40	0.38
Math	3	0.38	0.37	0.36	0.35	0.34	0.34	0.34	0.34	0.35	0.37	0.35
	4	0.42	0.37	0.35	0.34	0.34	0.34	0.34	0.35	0.35	0.36	0.35
	5	0.39	0.35	0.34	0.34	0.35	0.34	0.33	0.34	0.35	0.37	0.35
	6	0.41	0.37	0.36	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.36
	7	0.44	0.38	0.37	0.36	0.35	0.34	0.34	0.34	0.34	0.35	0.36
	8	0.47	0.40	0.38	0.37	0.36	0.35	0.35	0.34	0.34	0.34	0.37

Table 5.26. SEM by Deciles for NSCAS Scores—Spring Simulations

Content Area	Grade	Proficiency Score Distribution										Overall
		Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10	
ELA	3	0.58	0.45	0.40	0.38	0.37	0.36	0.36	0.35	0.36	0.41	0.40
	4	0.49	0.41	0.38	0.37	0.36	0.36	0.36	0.37	0.39	0.48	0.40

Content Area	Grade	Proficiency Score Distribution										Overall
		Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10	
	5	0.44	0.37	0.36	0.36	0.36	0.36	0.36	0.37	0.40	0.50	0.39
	6	0.47	0.39	0.36	0.34	0.34	0.34	0.35	0.36	0.39	0.50	0.38
	7	0.48	0.38	0.36	0.35	0.35	0.34	0.35	0.37	0.40	0.52	0.39
	8	0.46	0.37	0.35	0.34	0.33	0.34	0.35	0.37	0.41	0.53	0.39
Math	3	0.38	0.37	0.35	0.34	0.34	0.34	0.34	0.35	0.35	0.36	0.35
	4	0.42	0.37	0.35	0.34	0.35	0.34	0.35	0.35	0.35	0.36	0.36
	5	0.41	0.36	0.35	0.35	0.34	0.34	0.34	0.34	0.35	0.39	0.36
	6	0.39	0.36	0.35	0.34	0.35	0.34	0.34	0.34	0.34	0.35	0.35
	7	0.43	0.37	0.36	0.35	0.35	0.34	0.34	0.35	0.35	0.37	0.36
	8	0.46	0.39	0.37	0.36	0.35	0.35	0.35	0.35	0.35	0.36	0.37

Table 5.27. SEM by Deciles for NSCAS Scores—Spring Engine Evaluation

Content Area	Grade	Proficiency Score Distribution										Overall
		Decile 1	Decile 2	Decile 3	Decile 4	Decile 5	Decile 6	Decile 7	Decile 8	Decile 9	Decile 10	
ELA	3	0.43	0.39	0.37	0.36	0.35	0.35	0.36	0.36	0.38	0.43	0.38
	4	0.41	0.38	0.37	0.36	0.36	0.36	0.36	0.38	0.40	0.47	0.38
	5	0.40	0.36	0.36	0.36	0.36	0.36	0.36	0.37	0.38	0.44	0.38
	6	0.41	0.37	0.35	0.35	0.34	0.34	0.34	0.35	0.36	0.40	0.36
	7	0.41	0.37	0.36	0.35	0.35	0.35	0.35	0.35	0.36	0.41	0.37
	8	0.39	0.36	0.35	0.34	0.34	0.34	0.35	0.35	0.37	0.41	0.36
Math	3	0.37	0.36	0.34	0.34	0.34	0.34	0.35	0.35	0.36	0.41	0.36
	4	0.38	0.35	0.34	0.34	0.34	0.35	0.35	0.35	0.35	0.37	0.35
	5	0.37	0.34	0.34	0.34	0.34	0.33	0.34	0.34	0.34	0.38	0.35
	6	0.38	0.36	0.34	0.34	0.34	0.34	0.34	0.34	0.33	0.35	0.35
	7	0.40	0.36	0.36	0.35	0.34	0.34	0.34	0.34	0.34	0.35	0.35
	8	0.42	0.38	0.37	0.36	0.35	0.34	0.34	0.34	0.34	0.35	0.36

5.3. Engine Simulations: Science Field Test

The Spring 2023 science assessments are operational tests, following the Spring 2022 operational field tests (i.e., all items were re-calibrated following the 2022 administration). The number of items and points possible are reported in Table 2.2. Spring 2023 science tests can be summarized as:

- English online forms
 - Each grade has 20 different forms, but the operational items are the same across forms. The number of total items is presented in Table 2.2.
 - Each form of grade 5 has 9 sets (with 7 operational sets across forms and 2 field-test sets). Each form of grade 8 has 9 sets (with 6 operational sets across forms and 3 field-test sets). Each field-test set includes 4 to 8 items.
 - The overall test score is based on 31 operational items (worth 33 points) in grade 5 and 30 operational items (worth 33 points) in grade 8.

- The paper-pencil forms contain operational sets only, including field-test items associated with the operational sets. They include 33 items for both grades 5 and 8.
- Paper-pencil and Spanish forms
 - Each grade has 7 operational sets.
 - The overall test score is based on 31 operational items (worth 33 points) in grade 5 and 30 operational items (worth 33 points) in grade 8.

A simulation study and an engine evaluation check were conducted to provide evidence that the NWEA adaptive constraint-based engine (Cadabra) can properly administer the fixed forms as intended for the NSCAS science assessment for grades 5 and 8. The engine administered the fixed forms as intended; that is, prompts within a task were administered in a pre-specified fixed order, while operational tasks were ordered randomly, followed by field-test tasks.

Because science assessments are fixed forms with a small number of operational items, simulation and engine evaluations focused on whether each form was delivered to a representative sample of Nebraska students.

A total of 20,000 students per grade were included in the simulation study sample. The true values of student ability (theta, or θ) were drawn from a normal distribution with a mean of 0.0 and a standard deviation of 1.0. The student sample was simulated to have similar demographic characteristics to Nebraska's general student population based on the roster file, as shown in Table 5.28.

Table 5.29 through Table 5.32 present the number and percentage of simulated students who received each form by gender and ethnicity for grades 5 and 8, respectively. Each form was delivered to a representative sample of Nebraska students, demonstrating that the proportions set in the engine population exposure control were representative of the Nebraska general student population in terms of gender and ethnicity. Thus, it can be reasonably assumed that each field-test task and its prompts were also delivered to a representative sample of Nebraska students. These results suggest that the population exposure control function of the adaptive engine works well.

Table 5.28. General Population Demographic Distribution

Grade	Gender				Ethnicity								Total N		
	Female		Male		American Indian		Asian		Black		Hispanic			White	
	N	%	N	%	N	%	N	%	N	%	N	%		N	%
Nebraska General Population															
5	11,162	48.4	11,884	51.6	306	1.3	740	3.2	1,455	6.3	4,668	20.3	14,703	63.8	23,046
8	11,737	48.6	12,406	51.4	316	1.3	660	2.7	1,595	6.6	5,167	21.4	15,310	63.4	24,143
Simulation Student Sample															
5	9,688	48.4	10,312	51.6	256	1.3	577	2.9	1,231	6.2	3,958	19.8	13,127	65.6	20,000
8	9,564	47.8	10,436	52.2	271	1.4	483	2.4	1,160	5.8	3,812	19.1	13,522	67.6	20,000

Table 5.29. Demographic Distribution by Form—Grade 5 (Simulation)

Form	Gender				Ethnicity										Total N
	Female		Male		American Indian		Asian		Black		Hispanic		White		
	N	%	N	%	N	%	N	%	N	%	N	%	N	%	
(All)	9,688	48.4	10,312	51.6	256	1.3	577	2.9	1,231	6.2	3,958	19.8	13,127	65.6	20,000
A	493	49.2	510	50.8	12	1.2	30	3.0	61	6.1	204	20.3	653	65.1	1,003
B	484	48.9	505	51.1	4	0.4	25	2.5	58	5.9	203	20.5	665	67.2	989
C	502	50.3	497	49.7	12	1.2	30	3.0	59	5.9	211	21.1	650	65.1	999
D	513	50.7	499	49.3	13	1.3	23	2.3	65	6.4	205	20.3	660	65.2	1,012
E	470	48.0	510	52.0	10	1.0	30	3.1	62	6.3	177	18.1	655	66.8	980
F	485	49.3	499	50.7	10	1.0	24	2.4	65	6.6	196	19.9	643	65.3	984
G	502	50.0	502	50.0	11	1.1	27	2.7	66	6.6	202	20.1	660	65.7	1,004
H	475	47.7	521	52.3	10	1.0	28	2.8	72	7.2	190	19.1	663	66.6	996

Form	Gender				Ethnicity										Total N
	Female		Male		American Indian		Asian		Black		Hispanic		White		
	N	%	N	%	N	%	N	%	N	%	N	%	N	%	
I	462	47.2	517	52.8	7	0.7	32	3.3	55	5.6	199	20.3	646	66.0	979
J	489	47.2	546	52.8	13	1.3	33	3.2	69	6.7	203	19.6	678	65.5	1,035
K	492	48.6	521	51.4	22	2.2	30	3.0	62	6.1	204	20.1	640	63.2	1,013
L	493	49.5	503	50.5	17	1.7	25	2.5	55	5.5	202	20.3	666	66.9	996
M	475	48.0	514	52.0	13	1.3	30	3.0	57	5.8	206	20.8	625	63.2	989
N	491	49.5	500	50.5	12	1.2	31	3.1	57	5.8	197	19.9	648	65.4	991
O	454	45.5	543	54.5	16	1.6	34	3.4	60	6.0	195	19.6	659	66.1	997
P	476	47.6	523	52.4	18	1.8	32	3.2	67	6.7	192	19.2	653	65.4	999
Q	458	45.5	549	54.5	16	1.6	24	2.4	62	6.2	193	19.2	677	67.2	1,007
R	469	48.4	501	51.6	13	1.3	27	2.8	57	5.9	176	18.1	657	67.7	970
S	484	47.9	527	52.1	11	1.1	31	3.1	57	5.6	203	20.1	672	66.5	1,011
T	521	49.8	525	50.2	16	1.5	31	3.0	65	6.2	200	19.1	657	62.8	1,046

Table 5.30. Demographic Distribution by Form—Grade 8 (Simulation)

Form	Gender				Ethnicity								Total N		
	Female		Male		American Indian		Asian		Black		Hispanic			White	
	N	%	N	%	N	%	N	%	N	%	N	%		N	%
(All)	9,564	47.8	10,436	52.2	271	1.4	483	2.4	1,160	5.8	3,812	19.1	13,522	67.6	20,000

Form	Gender				Ethnicity										Total N
	Female		Male		American Indian		Asian		Black		Hispanic		White		
	N	%	N	%	N	%	N	%	N	%	N	%	N	%	
A	479	48.4	511	51.6	11	1.1	22	2.2	46	4.6	190	19.2	685	69.2	990
B	477	47.5	527	52.5	10	1.0	27	2.7	66	6.6	189	18.8	677	67.4	1,004
C	504	50.2	499	49.8	16	1.6	23	2.3	59	5.9	165	16.5	701	69.9	1,003
D	476	46.4	549	53.6	16	1.6	22	2.1	59	5.8	198	19.3	688	67.1	1,025
E	457	47.5	506	52.5	13	1.3	21	2.2	55	5.7	183	19.0	655	68.0	963
F	528	50.7	514	49.3	17	1.6	29	2.8	66	6.3	228	21.9	662	63.5	1,042
G	456	45.6	543	54.4	16	1.6	24	2.4	58	5.8	186	18.6	686	68.7	999
H	471	47.6	519	52.4	14	1.4	25	2.5	55	5.6	187	18.9	678	68.5	990
I	486	49.3	499	50.7	14	1.4	24	2.4	62	6.3	184	18.7	675	68.5	985
J	445	45.8	526	54.2	12	1.2	22	2.3	53	5.5	188	19.4	671	69.1	971
K	470	47.3	523	52.7	11	1.1	27	2.7	55	5.5	200	20.1	675	68.0	993
L	516	50.8	499	49.2	12	1.2	23	2.3	51	5.0	191	18.8	701	69.1	1,015
M	493	48.6	521	51.4	14	1.4	23	2.3	56	5.5	195	19.2	686	67.7	1,014
N	486	48.3	520	51.7	10	1.0	27	2.7	67	6.7	191	19.0	674	67.0	1,006
O	460	46.3	534	53.7	9	0.9	23	2.3	61	6.1	196	19.7	659	66.3	994
P	427	43.3	559	56.7	7	0.7	23	2.3	57	5.8	171	17.3	684	69.4	986
Q	463	46.3	538	53.7	19	1.9	23	2.3	62	6.2	187	18.7	677	67.6	1,001

Form	Gender				Ethnicity								Total N		
	Female		Male		American Indian		Asian		Black		Hispanic			White	
	N	%	N	%	N	%	N	%	N	%	N	%		N	%
R S T	493	49.9	494	50.1	19	1.9	21	2.1	54	5.5	194	19.7	660	66.9	987
	474	47.9	516	52.1	17	1.7	24	2.4	50	5.1	193	19.5	668	67.5	990
	503	48.3	539	51.7	14	1.3	30	2.9	68	6.5	196	18.8	660	63.3	1,042

Table 5.31. Demographic Distribution by Form—Grade 5 (Engine Evaluation)

Form	Gender				Ethnicity										Total N
	Female		Male		American Indian		Asian		Black		Hispanic		White		
	N	%	N	%	N	%	N	%	N	%	N	%	N	%	
(All)	11,090	48.5	11,785	51.5	284	1.2	739	3.2	1,433	6.3	4,580	20.0	14,575	63.7	22,875
A	598	49.7	605	50.3	11	0.9	41	3.4	82	6.8	222	18.5	773	64.3	1,203
B	548	48.8	575	51.2	15	1.3	31	2.8	62	5.5	222	19.8	731	65.1	1,123
C	550	48.4	587	51.6	14	1.2	30	2.6	69	6.1	235	20.7	731	64.3	1,137
D	564	45.4	677	54.6	20	1.6	43	3.5	79	6.4	220	17.7	744	60.0	1,241
E	523	46.4	603	53.6	17	1.5	36	3.2	68	6.0	221	19.6	730	64.8	1,126
F	560	49.8	565	50.2	10	0.9	36	3.2	66	5.9	231	20.5	725	64.4	1,125
G	559	50.9	539	49.1	11	1.0	34	3.1	47	4.3	222	20.2	697	63.5	1,098
H	580	48.7	612	51.3	16	1.3	56	4.7	102	8.6	240	20.1	746	62.6	1,192
I	514	46.2	598	53.8	11	1.0	30	2.7	56	5.0	235	21.1	698	62.8	1,112
J	562	48.4	599	51.6	29	2.5	42	3.6	77	6.6	238	20.5	722	62.2	1,161
K	566	51.5	534	48.5	14	1.3	28	2.5	60	5.5	237	21.5	676	61.5	1,100
L	558	48.8	585	51.2	9	0.8	37	3.2	78	6.8	234	20.5	733	64.1	1,143
M	575	50.9	555	49.1	8	0.7	35	3.1	65	5.8	222	19.6	737	65.2	1,130
N	564	49.1	585	50.9	10	0.9	32	2.8	82	7.1	230	20.0	756	65.8	1,149
O	537	47.8	587	52.2	23	2.0	40	3.6	63	5.6	230	20.5	706	62.8	1,124
P	546	48.6	578	51.4	7	0.6	32	2.8	69	6.1	229	20.4	735	65.4	1,124

Form	Gender				Ethnicity								Total N		
	Female		Male		American Indian		Asian		Black		Hispanic			White	
	N	%	N	%	N	%	N	%	N	%	N	%		N	%
Q	557	47.7	611	52.3	16	1.4	38	3.3	84	7.2	235	20.1	753	64.5	1,168
R	542	46.5	624	53.5	22	1.9	49	4.2	87	7.5	226	19.4	747	64.1	1,166
S	545	48.9	570	51.1	12	1.1	29	2.6	68	6.1	223	20.0	711	63.8	1,115
T	535	47.7	587	52.3	8	0.7	39	3.5	69	6.1	225	20.1	716	63.8	1,122

Table 5.32. Demographic Distribution by Form—Grade 8 (Engine Evaluation)

Form	Gender				Ethnicity								Total N		
	Female		Male		American Indian		Asian		Black		Hispanic			White	
	N	%	N	%	N	%	N	%	N	%	N	%		N	%
(All)	11,581	48.7	12,209	51.3	293	1.2	660	2.8	1,533	6.4	5,038	21.2	15,088	63.4	23,790
A	581	49.7	588	50.3	7	0.6	31	2.7	69	5.9	255	21.8	752	64.3	1,169
B	603	50.1	600	49.9	19	1.6	46	3.8	80	6.7	245	20.4	775	64.4	1,203
C	563	48.2	604	51.8	14	1.2	28	2.4	65	5.6	255	21.9	747	64.0	1,167
D	602	49.1	623	50.9	19	1.6	42	3.4	98	8.0	257	21.0	791	64.6	1,225
E	595	48.5	633	51.5	20	1.6	38	3.1	89	7.2	248	20.2	803	65.4	1,228
F	587	48.2	632	51.8	14	1.1	36	3.0	97	8.0	256	21.0	776	63.7	1,219
G	557	45.9	656	54.1	22	1.8	37	3.1	91	7.5	246	20.3	782	64.5	1,213
H	595	49.5	608	50.5	13	1.1	36	3.0	85	7.1	246	20.4	782	65.0	1,203
I	608	51.3	578	48.7	21	1.8	32	2.7	81	6.8	255	21.5	754	63.6	1,186
J	600	50.9	579	49.1	14	1.2	33	2.8	78	6.6	252	21.4	761	64.5	1,179
K	560	49.2	578	50.8	14	1.2	23	2.0	62	5.4	253	22.2	712	62.6	1,138
L	557	48.4	594	51.6	14	1.2	28	2.4	68	5.9	245	21.3	727	63.2	1,151
M	553	46.9	626	53.1	11	0.9	32	2.7	77	6.5	246	20.9	761	64.5	1,179
N	568	48.0	616	52.0	19	1.6	37	3.1	89	7.5	254	21.5	742	62.7	1,184
O	572	49.7	579	50.3	9	0.8	31	2.7	59	5.1	247	21.5	726	63.1	1,151
P	585	50.0	584	50.0	17	1.5	29	2.5	64	5.5	255	21.8	729	62.4	1,169
Q	636	47.3	708	52.7	12	0.9	39	2.9	89	6.6	254	18.9	788	58.6	1,344

Form	Gender				Ethnicity										Total N
	Female		Male		American Indian		Asian		Black		Hispanic		White		
	N	%	N	%	N	%	N	%	N	%	N	%	N	%	
R S T	568	48.3	608	51.7	11	0.9	31	2.6	79	6.7	256	21.8	749	63.7	1,176
	521	46.0	612	54.0	11	1.0	23	2.0	54	4.8	257	22.7	701	61.9	1,133
	554	48.6	586	51.4	11	1.0	27	2.4	53	4.6	248	21.8	717	62.9	1,140

Section 6: Psychometric Analyses

Pre-equated item parameter estimates were used to score student responses and select the next items to administer for the adaptive portions of the NSCAS Growth ELA and mathematics assessments. After the testing window was closed, the following post-administration analyses were conducted for ELA, mathematics, and science. The purpose of conducting these analyses is to establish the psychometric quality of the items used in the assessments, which will bolster arguments regarding the validity of the interpretations and uses of the test scores.

- Classical item analyses
- Differential item functioning (DIF)
- Item response theory (IRT) calibration

6.1. Number of Students Included in the Analyses

Table 6.1 presents the number of students included in the post-administration analyses (i.e., classical analyses, DIF, and IRT calibration). As in previous technical reports since 2018, only online test-takers who attempted at least 10 operational items were included. The results from these students are referred to as the “analyses data.” It is typically ideal to use 100% of the student data, including both online and paper-pencil tests; however, NDE decided to use only online tests due to the goal of completing the standard setting by the end of July 2018 and because the number of paper-pencil test-takers was less than 100 for each grade.

Table 6.1. Number of Students Included in the Psychometric Analyses

Content Area	Grade	Test ID	N
ELA	3	TB-766	23,257
	4	TB-767	22,913
	5	TB-768	22,977
	6	TB-769	22,850
	7	TB-770	23,430
	8	TB-771	23,881
Mathematics	3	TB-772	23,198
	4	TB-773	22,837
	5	TB-774	22,918
	6	TB-775	22,778
	7	TB-776	23,349
	8	TB-777	23,781
Science	5	TB-778	22,890
	8	TB-779	23,796

6.2. Classical Item Analyses

This section summarizes the p values and item-total correlations for operational and field-test items. Appendix B: Summary of P Values by Item Type and Appendix C: Summary of Item-Total Correlations by Item Type provide the classical item-level statistics. Omit rates across all content areas and grades were close to 0, which is to be expected since students were required to answer each item before moving on to the next one. Additionally, item statistics obtained from less than 100 students were not included for analyses.

6.2.1. Item Difficulty (P Value)

Item difficulty is measured by a p value, which shows the proportion of students who answered an item correctly and is bounded by 0 and 1. Generally, a high p value indicates that an item is easy (i.e., a high proportion of students answered it correctly), whereas a low p value indicates that an item is hard. For example, a p value of 0.79 indicates that 79% of students answered the item correctly. For polytomous items, the p value is the average item score (i.e., the sum of student scores on an item divided by the total number of students who responded to the item) divided by the number of possible score points on the item.

Table 6.2 and Table 6.3 present the summary statistics for the p values across all operational and field-test items, respectively, including the number of items by p -value range (i.e., less than or equal to a p value of 0.1, 0.2, etc.). These data were calculated for items with and without a representative sample (i.e., field-test items vs. adaptive items, respectively). Items without a representative sample are those administered during the adaptive stage of the assessment, and the expected p value is typically between 0.4 and 0.6 for these items. Appendix B: Summary of P Values by Item Type provides the summary p -value statistics by item type.

Table 6.2. Summary of *P* Values—Operational Items

Content Area	Gr.	#Items	Mean	SD	Min.	Max.	#Items by <i>P</i> -Value Range									
							≤ 0.1	≤ 0.2	≤ 0.3	≤ 0.4	≤ 0.5	≤ 0.6	≤ 0.7	≤ 0.8	≤ 0.9	> 0.9
ELA	3	639	0.52	0.14	0.11	0.92	0	3	27	83	169	182	105	53	16	1
	4	543	0.58	0.14	0.20	0.98	0	0	3	46	133	139	126	63	23	10
	5	546	0.56	0.13	0.05	0.95	1	0	10	44	126	156	130	54	20	5
	6	586	0.54	0.14	0.13	0.96	0	5	18	60	146	160	109	61	25	2
	7	506	0.55	0.13	0.10	0.95	0	2	12	41	121	173	101	39	14	3
	8	582	0.57	0.14	0.08	0.99	1	2	5	51	142	163	123	60	21	14
Math	3	795	0.51	0.10	0.00	0.88	2	5	4	64	359	238	82	33	8	0
	4	558	0.50	0.08	0.00	0.75	1	1	5	39	253	208	39	12	0	0
	5	611	0.50	0.09	0.00	1.00	1	0	2	55	267	214	56	14	1	1
	6	896	0.49	0.10	0.00	1.00	2	1	21	129	377	281	64	15	4	2
	7	677	0.46	0.10	0.00	1.00	1	4	23	115	324	179	23	4	1	3
	8	576	0.45	0.09	0.00	0.75	3	12	27	85	290	146	11	2	0	0
Science	5	31	0.58	0.19	0.13	0.88	0	2	1	1	5	6	7	5	4	0
	8	30	0.57	0.19	0.16	0.83	0	1	2	3	4	7	5	6	2	0

Table 6.3. Summary of *P* Values—Field-Test Items

Content Area	Gr.	#Items	Mean	SD	Min.	Max.	#Items by <i>P</i> -Value Range									
							≤ 0.1	≤ 0.2	≤ 0.3	≤ 0.4	≤ 0.5	≤ 0.6	≤ 0.7	≤ 0.8	≤ 0.9	> 0.9
ELA	3	161	0.59	0.17	0.15	0.97	0	1	10	13	24	40	32	21	12	8
	4	131	0.57	0.18	0.02	0.94	2	1	5	13	22	33	23	19	12	1
	5	170	0.58	0.20	0.14	0.99	0	4	11	22	25	28	25	34	13	8
	6	141	0.52	0.18	0.05	0.92	2	4	12	16	31	27	27	13	8	1
	7	150	0.55	0.20	0.03	0.93	2	4	11	20	24	27	25	24	10	3
	8	191	0.51	0.19	0.00	0.92	5	6	18	27	37	34	29	24	9	2
Math	3	13	0.25	0.11	0.04	0.43	2	2	5	3	1	0	0	0	0	0
	4	3	0.18	0.07	0.10	0.22	0	1	2	0	0	0	0	0	0	0
	5	6	0.37	0.15	0.20	0.64	0	1	1	2	1	0	1	0	0	0
	6	32	0.30	0.21	0.07	0.74	4	8	10	3	1	0	4	2	0	0
	7	10	0.24	0.13	0.04	0.48	1	3	3	2	1	0	0	0	0	0
	8	5	0.26	0.14	0.13	0.51	0	1	3	0	0	1	0	0	0	0
Science	5	119	0.53	0.17	0.02	0.96	1	3	5	12	29	30	25	8	4	2
	8	134	0.49	0.16	0.10	0.82	1	5	11	29	22	33	18	12	3	0

6.2.2. Item Discrimination (*Item-Total Correlation*)

Item-total correlation describes the relationship between performance on a specific item and performance on the entire test based on the student's overall test score. Students who do well on a test are expected to select the right answer to any given item, and students who do poorly are expected to select the wrong answer. This means that for a highly discriminating item, students who get the item correct will have a higher average test score than students who get the item incorrect. The item-total correlation coefficient ranges between -1.0 and +1.0. An item with a high positive item-total correlation discriminates between low-performing and high-performing students better than an item with an item-total correlation near zero. A negative item-total correlation indicates that lower-performing students did better on that item than higher-performing students. However, a very difficult item (or a very easy item) would have little variance in student responses, meaning most students respond incorrectly (or correctly). The resulting item-total correlation is typically low since both groups have the same score.

Table 6.4 and Table 6.5 present the summary statistics for the item-total correlations across all operational and field-test items, respectively. Appendix C: Summary of Item-Total Correlations by Item Type provides the results by item type. Instead of using the number-correct score, the estimated final theta score was used to compute the item-total correlations because number-correct scores would not provide much insight into student performance on an adaptive test since, in theory, all students get 50% correct on an adaptive assessment.

Table 6.4. Summary of Item-Total Correlations—Operational Items

Content Area	Grade	#Items	Mean	SD	Min.	Max.	#Items by Item-Total Correlation Range						
							≤ 0.1	≤ 0.2	≤ 0.3	≤ 0.4	≤ 0.5	≤ 0.6	> 0.6
ELA	3	639	0.38	0.10	0.05	0.66	1	22	112	239	193	63	9
	4	543	0.38	0.09	0.13	0.67	0	19	95	227	148	47	7
	5	546	0.36	0.09	0.12	0.65	0	21	112	247	121	41	4
	6	586	0.36	0.09	0.08	0.65	1	23	128	253	140	35	6
	7	506	0.37	0.10	0.02	0.64	2	16	106	189	151	36	6
	8	582	0.36	0.10	0.04	0.66	3	24	129	233	149	37	7
Math	3	795	0.28	0.13	-0.07	1.00	73	124	238	219	101	32	8
	4	558	0.27	0.14	-0.08	1.00	75	100	176	128	49	25	5
	5	611	0.28	0.13	-0.08	0.72	58	97	195	164	62	27	8
	6	896	0.30	0.14	-1.00	1.00	62	94	287	296	114	33	10
	7	677	0.30	0.14	-0.89	0.85	47	60	192	256	90	29	3
	8	576	0.25	0.14	-0.11	1.00	72	111	189	139	43	19	3
Science	5	31	0.46	0.10	0.28	0.63	0	0	3	3	14	9	2
	8	30	0.44	0.07	0.29	0.55	0	0	1	9	13	7	0

Table 6.5. Summary of Item-Total Correlations—Field-Test Items

Content Area	Grade	#Items	Mean	SD	Min.	Max.	#Items by Item-Total Correlation Range						
							≤ 0.1	≤ 0.2	≤ 0.3	≤ 0.4	≤ 0.5	≤ 0.6	> 0.6
ELA	3	161	0.37	0.10	-0.01	0.58	3	5	31	63	47	12	0
	4	131	0.34	0.14	-0.22	0.67	10	10	26	43	32	8	2
	5	170	0.32	0.13	-0.06	0.58	14	11	43	53	37	12	0
	6	141	0.32	0.12	-0.09	0.60	6	11	38	53	28	4	1
	7	150	0.35	0.13	0.02	0.70	2	14	38	49	32	10	5
	8	191	0.32	0.14	-0.15	0.61	16	15	46	54	42	17	1
Math	3	13	0.35	0.12	0.20	0.57	0	0	6	1	5	1	0
	4	3	0.18	0.14	0.04	0.32	1	1	0	1	0	0	0
	5	6	0.40	0.19	0.14	0.65	0	1	1	2	0	1	1
	6	32	0.31	0.17	-0.24	0.50	3	5	5	8	10	1	0
	7	10	0.39	0.09	0.24	0.53	0	0	1	4	3	2	0
	8	5	0.19	0.17	-0.09	0.35	1	1	2	1	0	0	0
Science	5	119	0.38	0.13	-0.09	0.64	5	10	10	32	49	10	3
	8	134	0.38	0.12	0.08	0.64	3	9	16	39	44	22	1

6.2.3. Item Suppression

Table 6.6 and Table 6.7 present the flagging criteria for multiple-choice (MC) and non-MC operational items, respectively. Based on the item analysis conducted using the spring administration results and removing items with n-counts less than 100 (statistics for items with $N < 100$ are considered to be unstable), 517 MC items and 101 non-MC items were identified for content and psychometric review. There were no science operational items flagged.

Table 6.6. Flagging Criteria for MC Items

Flag Type	Criterion	Indication
Low item-total	< 0.20	Poorly discriminating item
High item-total for a distractor	> 0.05	Poorly discriminating item

Note. item-total = item-total correlation

Table 6.7. Flagging Criteria for Non-MC Items

Flag Type	Criterion
Low item-total	< 0.10
High item-total for a score of 0	> 0
Item-total for a score of 1 is less than item-total for a score of 0	score of 1 item-total $<$ score of 0 item-total
Low item-total for a score of 0	< 0.10
Item-total for a score of 2 is less than item-total for a score of 1	score of 2 item-total $<$ score of 1 item-total
Low student count for each score	< 0

Note. item-total = item-total correlation. All flags in this table indicate poor discrimination.

After the content and psychometric teams reviewed these flagged items, NWEA recommended suppressing three items from scoring and removing them from the item pool, as shown in Table 6.8. Following NDE approval, these suppressed items were not included for all subsequent analyses and score reporting. There was no suppression for science operational items.

Table 6.8. Items to Be Suppressed

Content Area	Grade	Item Code	Item Role ^a	Item Type	Standard (Indicator)	Max. #Points	NWEA Recommendations	
							2022–2023 Spring Scoring	2023–2024 Pool & Later
ELA	5	VR431908	OP	Choice-Single	LA.5.RP.1	1	Suppress	Remove from the pool
Math	4	VR463874	DO	Choice-Single	MA 4.1.1.a	1	Suppress	Remove from the pool
Math	8	VR468448	DO	Choice-Single	MA 8.1.1.b	1	Suppress	Remove from the pool

^a OP = operational; DO = diagnostic operational

6.3. Differential Item Functioning (DIF)

Differential item functioning (DIF) is a statistical procedure that flags items for potential bias. The fundamental measurement assumption of DIF is that the probability of a correct response to a test item is a function of the item's difficulty and the student's ability. This function is expected to remain invariant to other characteristics unrelated to ability, such as gender and ethnicity. Therefore, if two students with the same ability respond to the same item, they are assumed to have an equal probability of answering the item correctly. To test this assumption, responses to items by students sharing an aspect of a characteristic (e.g., gender) are compared with responses to the same items by other students who share a different aspect of the same characteristic (e.g., males vs. females). The group representing students in a specific demographic group is referred to as the *focal* group. The group comprised of students from outside this group is referred to as the *reference* group. Table 6.9 presents the focal and reference groups for the NSCAS DIF analyses.

Table 6.9. Focal and Reference Groups for Gender- and Ethnicity-Based DIF

Group Type	Focal Group	Reference Group
Gender	Female	Male
Ethnicity	Black or African American	White
	Hispanic	White
	Asian	White
	Two or More Races	White

When DIF is detected and the fundamental measurement assumption does not hold (i.e., students with the same ability in different groups of interest have different probabilities of correctly answering an item), the item is said to be functioning differently for the two groups. The presence of DIF in an item suggests that the item is functioning unexpectedly regarding the groups included in the comparison. The cause of the unexpected functioning is not revealed in a DIF analysis. It may be that item content is inadvertently providing an advantage or disadvantage to members of one of the two groups. Content experts who have special knowledge of the groups involved can often identify a cause of this type. DIF may also result from differential instruction closely associated with group membership.

Because fairness is a fundamental validity issue, it is essential that items be reviewed and assessed for DIF. Many methods for assessing DIF have been used and compared in conventional paper-pencil non-adaptive tests; however, DIF detection may be more important for a CAT than it is for traditional paper-pencil non-adaptive tests for two reasons (Zwick et al., 1994): First, items with DIF may be more consequential for the examinees because fewer items are administered in a CAT. Second, several potential sources of DIF may be introduced, such as differential computer familiarity, facility, and anxiety. The difficulty of DIF analysis in a CAT is introduced by the fact that different sets of items are administered to different examinees. Therefore, the logistic regression (LR) procedure was applied to ELA and mathematics items that were administered in this CAT, while the Mantel-Haenszel (MH) procedure was used for science items that were administered as a fixed form.

6.3.1. Logistic Regression (LR) DIF Method

The LR DIF procedure models item responses (for both dichotomous and polytomous items) as a function of group memberships, ability estimates, and their interaction. Testing for the presence of DIF based on logistic regression provides a model-based approach to identify

uniform and nonuniform DIF. DIF is classified as uniform if the effect is constant; that is, uniform DIF (UIDIF) exists when the difference in the probabilities of a correct answer for the two groups is the same at all ability levels. DIF is classified as nonuniform (NUIDIF) if the effect varies conditional on the ability level; that is, nonuniform DIF exists if the interaction between item-response function and group membership is disordinal.

The LR DIF procedure compares the following three models (Fu & Monfils, 2016; Swaminathan & Rogers, 1990; Zumbo, 1999):

$$\text{Model 1: } \text{logit}(P) = \beta_0 + \beta_1 X + \beta_2 E$$

$$\text{Model 2: } \text{logit}(P) = \beta_0 + \beta_1 X + \beta_2 G + \beta_3 E$$

$$\text{Model 3: } \text{logit}(P) = \beta_0 + \beta_1 X + \beta_2 G + \beta_3 XG + \beta_4 E$$

where:

- P is the probability of a test taker answering an item incorrectly (for a dichotomous item) and the probability of getting an item score or lower (for a polytomous item).
- X is the criterion variable.
- G is the group membership.
- E is a vector, including additional explanatory variables.
- β are the associated regression parameters for model k .

For both dichotomous and polytomous items, Models 1, 2, and 3 are also referred to as a no-DIF model, a uniform DIF model, and a nonuniform DIF model, respectively. The group estimates (β_2) are related to uniform DIF, and the interaction estimates (β_3) are associated with nonuniform DIF. Note that for a dichotomously scored item, the target probability that the LR estimates is the probability of answering an item incorrectly, which is different from the probability of answering an item correctly that many people may be accustomed to. Similarly, the target probability in the regression model for a polytomously scored item is the probability of obtaining an item score or below, to be consistent with that for a dichotomously scored item.

The item shows DIF if the modeled fit statistic is improved when group and interaction are added to the model, in order. To test the presence of nonuniform DIF, Model 2 and Model 3 are compared, using the likelihood ratio test with 1 degree of freedom (df) in chi-square distribution:

$$\chi^2 = [-2 \ln L(\text{Model2})] - [-2 \ln L(\text{Model3})].$$

Similarly, to test the presence of uniform DIF, Model 1 and Model 2 are compared, using the likelihood ratio test with 1 df:

$$\chi^2 = [-2 \ln L(\text{Model1})] - [-2 \ln L(\text{Model2})].$$

To test overall DIF (uniform DIF or nonuniform DIF), Model 1 and Model 3 are compared, using the likelihood ratio test with 2 df:

$$\chi^2 = [-2 \ln L(\text{Model1})] - [-2 \ln L(\text{Model3})].$$

The effect size is also used to avoid practically trivial but statistically significant results (French & Miller, 1996). Effect size is indicated by the difference of the Nagelkerke R^2 between two models (Gómez-Benito et al., 2009). Table 6.10 presents the DIF classification rules for the LR

DIF procedure used for NSCAS. These rules were confirmed to be consistent with the MH DIF classification rule for dichotomous items used by Educational Testing Service (ETS) (Fu & Monfils, 2016).

Table 6.10. LR DIF Categories

DIF Category	Level of DIF	Definition
A	Negligible	χ^2 test is not significant at 0.05 level or $\Delta R^2 < 0.035$.
B	Moderate	χ^2 test is significant at 0.05 level and $0.035 \leq \Delta R^2 < 0.070$.
C	Strong	χ^2 test is significant at 0.05 level and $\Delta R^2 \geq 0.070$.

Note. ΔR^2 is the Nagelkerke R^2 difference between two models.

6.3.2. Mantel-Haenszel (MH) DIF Methods

The Mantel-Haenszel (MH) procedure was used to detect DIF for dichotomous items (Holland & Thayer, 1988), and the standardized mean difference (SMD) analysis, developed as an extension of the MH procedure, was used to detect DIF for polytomous items (Dorans & Schmitt, 1991; Zwick et al., 1993). The MH method has been widely used in educational measurement due to its easy implementation in testing programs. The procedure compares the ratio of the probabilities of two groups of students (i.e., the focal and reference groups) answering an item correctly across all score levels. The obtained estimate is known as the odds ratio, which is computed as follows:

$$\alpha_{MH} = \frac{\left(\sum_m \frac{R_{rm} W_{fm}}{N_m} \right)}{\left(\sum_m \frac{R_{fm} W_{rm}}{N_m} \right)}$$

where:

- R_{rm} is the number of students in the reference group at ability level m answering the item correctly.
- W_{fm} is the number of students in the focal group at ability level m answering the item incorrectly.
- R_{fm} is the number of students in the focal group at ability level m answering the item correctly.
- W_{rm} is the number of students in the reference group at ability level m answering the item incorrectly.
- N_m is the total number of students at ability level m .

This value can then be used as follows (Holland & Thayer, 1988):

$$MH\ D - DIF = -2.35 \ln(\alpha_{MH})$$

The MH chi-square statistic used to classify items into the three ETS DIF categories is as follows:

$$MH\ CHISQ = \frac{\left(\left| \sum_m R_{rm} - \sum_m E(R_{rm}) \right| - \frac{1}{2} \right)^2}{\sum_m Var(R_{rm})}$$

where:

- $E(R_{rm}) = \frac{N_{rm}R_{Nm}}{N_m}$, $Var(R_{rm}) = \frac{N_{rm}N_{fm}R_{Nm}W_{Nm}}{N_m^2(N_{m-1})}$
- N_{rm} and N_{fm} are the number of students in the reference and focal groups, respectively.
- R_{Nm} and W_{Nm} are the number of students who answered the item correctly and incorrectly, respectively.

Standardized mean difference (SMD) for polytomous items compares the item performance of two subpopulations, adjusting for differences in the distributions of the two subpopulations. The SMD statistic can be divided by the total standard deviation to obtain a measure of the effect size. A negative value of the standardized mean difference shows that the item is more difficult for the focal group, whereas a positive value indicates that it is more difficult for the reference group. The standardized mean difference used for polytomous items is defined as:

$$SMD = \sum p_{FK}m_{FK} - \sum p_{RK}m_{RK}$$

where:

- p_{FK} is the proportion of the focal group students at the K_{th} level of the matching criterion variable.
- m_{FK} is the mean score for the focal group at the K_{th} level of the matching criterion variable.
- p_{RK} is the proportion of the reference group students at the K_{th} level of the matching criterion variable.
- m_{RK} is the mean item score for the reference group at the K_{th} level of the matching criterion variable.

The SMD is divided by the total item group standard deviation to get a measure of the effect size.

Table 6.11 and Table 6.12 present the ETS DIF categories for classifying the DIF results. The ETS method of categorizing DIF allows items exhibiting negligible DIF (Category A) to be differentiated from those exhibiting moderate DIF (Category B) and strong DIF (Category C). Categories B and C have a further breakdown as “+” (DIF is in favor of the focal group) or “-” (DIF is in favor of the reference group).

Table 6.11. MH DIF Categories for Dichotomous Items

DIF Category	Level of DIF	Definition
A	Negligible	$MH \chi^2$ test is not significant at 0.05 level or $ MH \text{ D-DIF} < 1.0$.
B	Moderate	$MH \chi^2$ test is not significant at 0.05 level and $1.0 \leq MH \text{ D-DIF} < 1.5$.
C	Strong	$MH \chi^2$ test is not significant at 0.05 level and $ MH \text{ D-DIF} \geq 1.5$.

Note. $|MH \text{ D-DIF}|$ = absolute value of the Mantel-Haenszel delta difference

Table 6.12. MH DIF Categories for Polytomous Items

DIF Category	Level of DIF	Definition
A	Negligible	$MH \chi^2$ test is not significant at 0.05 level or $ SMD/SD \leq 0.17$.
B	Moderate	$MH \chi^2$ test is not significant at 0.05 level and $0.17 < SMD/SD \leq 0.25$.
C	Strong	$MH \chi^2$ test is not significant at 0.05 level and $ SMD/SD > 0.25$.

Note. SMD = standardized mean difference; SD = standard deviation

6.3.3. DIF Results

“Male” was the reference group for gender, and “white” was the reference group for ethnicity. DIF was not conducted if the sample size for either group was less than 100, which is reduced from 250 due to the increased number of field-test items. The “+” sign next to the DIF category indicates that the item is in favor of the reference group, and the “-” sign indicates that the item is in favor of the focal group.

Table 6.13 and Table 6.14 present the number of field-test items assigned to each LR DIF category for DIF and UIDIF, respectively, for ELA and mathematics. Considering that the Rasch model is applied (i.e., the same slope is assumed for all items), NUIDIF results are not reported.

Beginning in Spring 2021, item exposure is being controlled by an adaptive engine feature that assigns a weight to an item based on the number of times the item is seen by students. This feature resulted in increased item-pool usage, which is one of the desired properties that adaptive testing can achieve. However, it reduced the number of operational items meeting the minimum student counts required for DIF analyses because all operational items were selected adaptively, while field-test item distribution was controlled to meet required students counts and to be administered across demographics. Thus, the DIF results for field-test items in ELA and mathematics are reported. Table 6.15 presents the number of items assigned to each MH DIF category for science operational and field-test items, respectively. As shown in the tables, most items were categorized as DIF Category A (negligible DIF).

Table 6.13. LR DIF Results—Field-Test Items (ELA/Mathematics)

Content Area	Grade	Focal Group	#Items by DIF Category						C+	C-
			Total	A	B	B+	B-	C		
ELA	3	Female	161	161	--	--	--	--	--	--
		Black or African American	20	20	--	--	--	--	--	--
		Hispanic	161	157	3	--	1	--	--	--
		Asian	--	--	--	--	--	--	--	--
		Two or More Races	--	--	--	--	--	--	--	--
	4	Female	131	131	--	--	--	--	--	--
		Black or African American	25	25	--	--	--	--	--	--
		Hispanic	131	130	--	--	--	1	--	--
		Asian	--	--	--	--	--	--	--	--
		Two or More Races	--	--	--	--	--	--	--	--
	5	Female	170	168	1	1	--	--	--	--
		Black or African American	6	6	--	--	--	--	--	--
		Hispanic	170	166	3	--	--	1	--	--
		Asian	--	--	--	--	--	--	--	--
		Two or More Races	--	--	--	--	--	--	--	--
	6	Female	141	141	--	--	--	--	--	--
		Black or African American	5	5	--	--	--	--	--	--
		Hispanic	141	140	1	--	--	--	--	--
		Asian	--	--	--	--	--	--	--	--
		Two or More Races	--	--	--	--	--	--	--	--
	7	Female	150	148	2	--	--	--	--	--

Content Area	Grade	Focal Group	#Items by DIF Category							
			Total	A	B	B+	B-	C	C+	C-
		Black or African American	8	8	--	--	--	--	--	--
		Hispanic	150	146	3	--	--	1	--	--
		Asian	--	--	--	--	--	--	--	--
		Two or More Races	--	--	--	--	--	--	--	--
		Female	189	188	1	--	--	--	--	--
	8	Black or African American	5	5	--	--	--	--	--	--
		Hispanic	189	189	--	--	--	--	--	--
		Asian	--	--	--	--	--	--	--	--
		Two or More Races	--	--	--	--	--	--	--	--
		Math	3	Female	13	13	--	--	--	--
Black or African American	9			9	--	--	--	--	--	--
Hispanic	13			12	1	--	--	--	--	--
Asian	--			--	--	--	--	--	--	--
Two or More Races	--			--	--	--	--	--	--	--
4	Female		3	3	--	--	--	--	--	--
	Black or African American		3	3	--	--	--	--	--	--
	Hispanic		3	3	--	--	--	--	--	--
	Asian		--	--	--	--	--	--	--	--
	Two or More Races		--	--	--	--	--	--	--	--
5	Female		6	6	--	--	--	--	--	--
	Black or African American		6	6	--	--	--	--	--	--
	Hispanic		6	6	--	--	--	--	--	--
	Asian		--	--	--	--	--	--	--	--
	Two or More Races		--	--	--	--	--	--	--	--
6	Female		32	32	--	--	--	--	--	--
	Black or African American		2	2	--	--	--	--	--	--
	Hispanic		32	29	2	--	1	--	--	--
	Asian		--	--	--	--	--	--	--	--
	Two or More Races		--	--	--	--	--	--	--	--
7	Female		10	10	--	--	--	--	--	--
	Black or African American		10	10	--	--	--	--	--	--
	Hispanic		10	9	1	--	--	--	--	--
	Asian		--	--	--	--	--	--	--	--
	Two or More Races		--	--	--	--	--	--	--	--
8	Female		5	5	--	--	--	--	--	--
	Black or African American		5	5	--	--	--	--	--	--
	Hispanic		5	5	--	--	--	--	--	--
	Asian		--	--	--	--	--	--	--	--
	Two or More Races	--	--	--	--	--	--	--	--	

Table 6.14. LR UIDIF Results—Field-Test Items (ELA/Mathematics)

Content Area	Grade	Focal Group	#Items by DIF Category					
			Total	A	B+	B-	C+	C-
ELA	3	Female	161	161	--	--	--	--
		Black or African American	20	20	--	--	--	--
		Hispanic	161	158	--	3	--	--
		Asian	--	--	--	--	--	--
		Two or More Races	--	--	--	--	--	--
	4	Female	131	131	--	--	--	--
		Black or African American	25	25	--	--	--	--
		Hispanic	131	130	--	--	--	1
		Asian	--	--	--	--	--	--
		Two or More Races	--	--	--	--	--	--
	5	Female	170	169	1	--	--	--
		Black or African American	6	6	--	--	--	--
		Hispanic	170	169	--	1	--	--
		Asian	--	--	--	--	--	--
		Two or More Races	--	--	--	--	--	--
	6	Female	141	141	--	--	--	--
		Black or African American	5	5	--	--	--	--
		Hispanic	141	140	--	1	--	--
		Asian	--	--	--	--	--	--
		Two or More Races	--	--	--	--	--	--
	7	Female	150	149	--	1	--	--
		Black or African American	8	8	--	--	--	--
		Hispanic	150	149	--	--	--	1
		Asian	--	--	--	--	--	--
		Two or More Races	--	--	--	--	--	--
	8	Female	189	188	1	--	--	--
		Black or African American	5	5	--	--	--	--
		Hispanic	189	189	--	--	--	--
		Asian	--	--	--	--	--	--
		Two or More Races	--	--	--	--	--	--
Math	3	Female	13	13	--	--	--	--
		Black or African American	9	9	--	--	--	--
		Hispanic	13	13	--	--	--	--
		Asian	--	--	--	--	--	--
		Two or More Races	--	--	--	--	--	--
	4	Female	3	3	--	--	--	--
		Black or African American	3	3	--	--	--	--
		Hispanic	3	3	--	--	--	--
		Asian	--	--	--	--	--	--
		Two or More Races	--	--	--	--	--	--
	5	Female	6	6	--	--	--	--
		Black or African American	6	6	--	--	--	--
		Hispanic	6	6	--	--	--	--
		Asian	--	--	--	--	--	--
		Two or More Races	--	--	--	--	--	--
	6	Female	32	32	--	--	--	--
		Black or African American	2	2	--	--	--	--
		Hispanic	32	30	--	2	--	--
		Asian	--	--	--	--	--	--
		Two or More Races	--	--	--	--	--	--
	7	Female	10	10	--	--	--	--

Content Area	Grade	Focal Group	#Items by DIF Category					
			Total	A	B+	B-	C+	C-
		Black or African American	10	10	--	--	--	--
		Hispanic	10	10	--	--	--	--
		Asian	--	--	--	--	--	--
		Two or More Races	--	--	--	--	--	--
	8	Female	5	5	--	--	--	--
		Black or African American	5	5	--	--	--	--
		Hispanic	5	5	--	--	--	--
		Asian	--	--	--	--	--	--
		Two or More Races	--	--	--	--	--	--

Table 6.15. MH DIF Results—Field-Test Items (Science)

Content Area	Grade	Focal Group	#Items by DIF Category					
			Total	A	B+	B-	C+	C-
Science	5	Female	119	108	5	4	--	2
		Black or African American	8	6	--	2	--	--
		Hispanic	119	111	2	5	1	--
		Asian	2	2	--	--	--	--
		Two or More Races	2	2	--	--	--	--
	8	Female	134	125	3	5	--	1
		Black or African American	3	3	--	--	--	--
		Hispanic	134	125	--	7	1	1
		Asian	3	3	--	--	--	--
		Two or More Races	3	3	--	--	--	--

6.4. IRT Calibration

The Rasch model (Rasch, 1960, 1980; Wright, 1977) for dichotomous items and the partial-credit model (PCM; Masters, 1982) for polytomous items were used to calibrate items and create the NSCAS scale. For all content areas, item parameter estimations were implemented using WINSTEPS 3.91.0.0 (Linacre, 2015) that used joint maximum likelihood estimation (MLE), as described by Wright (1977) and Masters (1982). The Rasch model has had a long-standing presence in applied testing programs and was the methodology used to calibrate the previous Nebraska State Accountability (NeSA) items.

Under the Rasch model, the probability of a student with ability θ responding correctly to item i is as follows, where θ_j and b_i are the person and item parameters, respectively:

$$P(u_{ij} = 1 \mid \theta_j, b_i) = \frac{e^{(\theta_j - b_i)}}{1 + e^{(\theta_j - b_i)}}$$

Under the PCM, the probability of a student with ability θ having a score at the k th level of item i is:

$$P(u_{ij} = k \mid \theta_i) = \frac{e^{[\sum_{u=1}^k (\theta_j - b_i + d_{iu})]}}{\sum_{v=1}^m e^{[\sum_{u=1}^k (\theta_j - b_i + d_{iu})]}}$$

where k is the score on the item, m_i is the total number of score categories for the item, d_{iu} is the threshold parameter for the threshold between scores u and $u - 1$, and θ_j and b_i are the person and item parameters, respectively.

6.4.1. Summary of IRT Item Statistics

Table 6.16 and Table 6.17 present the summary of IRT item statistics across all operational and field-test items, respectively. The mean of the operational item parameters increases by grade for ELA and mathematics, as can be expected for vertical scales.

Table 6.16. Summary of IRT Item Statistics—Operational Items

Content Area	Grade	#Items	#Parameters	Mean	SD	Min.	Max.	Range (Max.–Min.)
ELA	3	625	712	-0.68	1.16	-3.52	3.71	7.23
	4	541	647	-0.52	1.21	-7.57	3.49	11.06
	5	538	644	-0.31	2.66	-42.39	43.61	86.00
	6	568	682	0.06	2.76	-46.21	45.34	91.56
	7	493	591	0.12	1.09	-2.72	4.84	7.56
	8	566	695	0.26	1.20	-5.65	5.64	11.29
Math	3	791	859	-0.62	1.43	-4.73	6.30	11.03
	4	557	631	0.30	1.34	-3.18	5.08	8.26
	5	610	684	0.34	1.35	-4.15	5.26	9.41
	6	895	982	0.73	1.44	-3.65	5.36	9.01
	7	677	752	1.24	1.41	-2.94	6.02	8.96
	8	572	642	1.49	1.42	-2.40	5.54	7.95
Science	5	31	33	0.30	1.29	-2.06	3.64	5.70
	8	30	33	-0.75	1.00	-2.15	1.69	3.84

Table 6.17. Summary of IRT Item Statistics—Field-Test Items

Content Area	Grade	#Items	#Parameters	Mean	SD	Min.	Max.	Range (Max.–Min.)
ELA	3	161	184	-0.90	1.12	-4.53	1.87	6.40
	4	131	156	-0.43	1.17	-3.32	4.18	7.50
	5	170	197	-0.31	1.22	-4.39	3.45	7.83
	6	141	168	0.18	1.11	-2.58	3.16	5.73
	7	150	189	0.12	4.64	-41.10	45.78	86.87
	8	189	226	0.53	1.22	-2.74	5.04	7.78
Math	3	13	17	1.60	1.35	-0.53	4.39	4.92
	4	3	3	2.59	0.66	2.19	3.36	1.17
	5	6	9	1.29	1.42	-1.16	3.04	4.21
	6	32	41	2.05	1.42	-0.83	4.57	5.40
	7	10	12	2.57	1.50	0.23	5.14	4.92
	8	5	7	2.30	2.10	-2.09	4.06	6.15
Science	5	119	148	0.60	1.15	-2.76	5.46	8.22
	8	134	170	-0.29	0.99	-2.40	2.45	4.85

6.5. Scaling

For science, the scaling constants were updated following the 2023 standard validation. For ELA and mathematics, scaling constants were set in 2018 without anchoring cut scores. After constructing the vertical scales for ELA and mathematics, descriptive statistics of student scale scores were examined to determine the following scaling constants of slope and intercept:

- A slope of $66.6/\sigma_{G5}$ (i.e., slope = 72.47244) and an intercept of 2500 for ELA
- A slope of $66.6/\sigma_{G5}$ (i.e., slope = 54.92622) and an intercept of 1200 for mathematics

where σ_{G5} is the standard deviation of the grade 5 theta score.

The theta estimate, θ , and associated θ -CSEM of students were then expressed on the NSCAS reporting scale by applying the linear transformation, slope and intercept (A and B, respectively), as follows:

$$\begin{aligned}SS &= (\theta \times A) + B \\SSCSEM &= \theta\text{-CSEM} \times A.\end{aligned}$$

θ -CSEM is defined as the reciprocal of the square root of the test information function and can be estimated across all points of the ability continuum (Hambleton & Swaminathan, 1985):

$$\theta\text{-CSEM} = CSEM(\theta_i) = \frac{1}{\sqrt{I(\theta_i)}}$$

where $I(\theta)$ is the test information function, as a sum of the item information function, obtained as:

$$I(\theta_i) = \sum_j \frac{p'_{ij}(\theta_i)^2}{p_{ij}(\theta_i) q_{ij}(\theta_i)}$$

where $p'_{ij}(\theta_i)$ is the derivative of $p_{ij}(\theta_i)$ and $q_{ij}(\theta_i) = 1 - p_{ij}(\theta_i)$. Once the linear transformation was applied, the scaled scores and associated CSEMs were rounded to an integer value. There was no adjustment made around cut scores or the scale score CSEM (SSCSEM). Final adjustments were made to scale scores that fell outside of the highest obtainable scale score (HOSS) or the lowest obtainable scale score (LOSS).

In setting the HOSS for ELA and mathematics, the following guidelines were considered. In setting the LOSS, similar guidelines were considered.

1. The HOSS must increase as the grade increases for tests on a vertical scale.
2. The HOSS should be high enough that it does not cause an unnecessary “pile-up” of scale scores at the HOSS, targeting less than 1%.
3. The HOSS should be low enough that $SSCSEM(HOSS) < 10 \times \text{Min}(SSCSEM)$.
4. The HOSS may be high enough that $SSCSEM(\text{Penultimate HOSS}) < 5 \times \text{Min}(SSCSEM)$.
5. The HOSS gap should not be too small, as a future test form may be slightly more difficult. It is also important that the gap is not too large, as that will tend to impact the mean of the distribution for cases with many perfect scores.

6. The gaps should change smoothly over score points, and the HOSS gap should transition smoothly across grades. It is more difficult, and less important, to keep the gaps smooth over score points and grades than it is to keep the SSCSEM values smooth over score points and SSCSEM (HOSS) transitions smooth across grade levels.

Based on these guidelines, the LOSS and HOSS presented in Table 6.18 were used. To be consistent with ELA and mathematics score ranges, the LOSS of science was changed from 1 to 0. This did not, however, change actual scores in that a score of 0 was assigned to students who attempted 0 items and a score of 1 was assigned to students who attempted 1–9 operational items. However, this change did make the communication consistent: the LOSS of each grade was used for students with 0 items attempted, a score of one point higher than LOSS was used for students who attempted 1–9 operational items, and a score of two points higher than LOSS was used for students who attempted 10 or more operational items.

Table 6.18. Score Range (LOSS and HOSS) and Assigned Score

Content Area	Grade	LOSS	HOSS	Assigned Score for Students with 0 OP Items Attempted	Assigned Score for Students with 1–9 OP Items Attempted	Lowest Calculated Score for Students with 10 or more OP Items Attempted
ELA	3	2220	2840	2220	2221	2222
	4	2250	2850	2250	2251	2252
	5	2280	2860	2280	2281	2282
	6	2290	2870	2290	2291	2292
	7	2300	2880	2300	2301	2302
	8	2310	2890	2310	2311	2312
Math	3	1000	1470	1000	1001	1002
	4	1010	1500	1010	1011	1012
	5	1020	1510	1020	1021	1022
	6	1030	1530	1030	1031	1032
	7	1040	1540	1040	1041	1042
	8	1050	1550	1050	1051	1052
Science	5	3000	3250	3000	3001	3002
	8	3000	3250	3000	3001	3002

Table 6.19 summarizes the cut-score implementation, or the conversion of student ability (theta) to scale scores that were used for scoring, which were updated based on the 2023 standard setting (see Section 8 for details). Specifically, the table presents the calculations of the slopes and intercepts for all grades of the scale score conversions, including the cut scores set during standard setting.

Table 6.19. Conversion of Theta to Scale Scores

Content Area	Grade	Scale Score Ranges by Achievement Levels			Cuts (Scale Scores)		Cuts (Theta)		Transformation Constants	
		Level 3: <i>Developing</i>	Level 2: <i>On Track</i>	Level 1: <i>Advanced</i>	Dev/ OT	OT/ Adv	Dev/ OT	OT/ Adv	Slope	Intercept
ELA	3	2220–2442	2443–2535	2536–2840	2443	2536	-0.791	0.503	72.4724	2500
	4	2250–2492	2493–2566	2567–2850	2493	2567	-0.096	0.921	72.4724	2500
	5	2280–2503	2504–2582	2583–2860	2504	2583	0.060	1.141	72.4724	2500
	6	2290–2517	2518–2593	2594–2870	2518	2594	0.242	1.303	72.4724	2500
	7	2300–2526	2527–2608	2609–2880	2527	2609	0.373	1.510	72.4724	2500
	8	2310–2523	2524–2623	2624–2890	2524	2624	0.326	1.707	72.4724	2500
Math	3	1000–1175	1176–1296	1297–1470	1176	1297	-0.440	1.770	54.9262	1200
	4	1010–1207	1208–1330	1331–1500	1208	1331	0.154	2.386	54.9262	1200
	5	1020–1206	1207–1319	1320–1510	1207	1320	0.136	2.179	54.9262	1200
	6	1030–1227	1228–1320	1321–1530	1228	1321	0.507	2.200	54.9262	1200
	7	1040–1211	1212–1313	1314–1540	1212	1314	0.219	2.084	54.9262	1200
	8	1050–1230	1231–1318	1319–1550	1231	1319	0.563	2.158	54.9262	1200
Science	5	3000–3099	3100–3149	3150–3250	3100	3150	-0.15	2.22	21.0971	3103.165
	8	3000–3099	3100–3149	3150–3250	3100	3150	-0.79	1.02	27.5346	3121.843

Section 7: Standard Setting

In 2023, NWEA contracted with ACS Ventures, LLC (ACS) to conduct a standard setting for the NSCAS for grades 3 through 8 in mathematics and ELA. NWEA also contracted ACS to conduct a standard validation for grades 5 and 8 in science.

7.1. ELA and Mathematics

7.1.1. Methodology

ACS used the procedures described within the Bookmark method of standard setting (Lewis et al., 2012) to guide panels of Nebraska educators through the process of recommending two cut scores to be used to distinguish the three achievement levels (i.e., *Developing*, *On Track*, *Advanced*):

1. The cut score that differentiates *On-Track* performance from *Developing* performance
2. The cut score that differentiates *Advanced* performance from *On-Track* performance

A key feature of the Bookmark method is presentation of the assessment items in an ordered item booklet (OIB). Specifically, the assessment content is presented in order of difficulty, starting with the easiest item and progressing through more and more difficult items. Expert panelists are instructed to review the OIB and identify the expected level of performance for a student who is just barely within the *On-Track* achievement level and then identify the expected level of performance for a student who is just barely within the *Advanced* achievement level.

7.1.2. Meeting Process

The standard setting study occurred over three days from July 25–July 28, 2023. The primary standard setting activities (large group orientation and training, panel-specific training, iterative judgmental process) were conducted during the first two days and half of the third. A subset of each panel was then asked to participate in an articulation meeting on the third afternoon to review the results across grades and subject areas.

The first part of the meeting served as an introduction to the general standard setting process. It began with a large-group general session that included a welcome and introductions from NDE. Then, ACS lead facilitators provided a high-level orientation and training on the standard setting process and methodology that was to be followed. The overview also included a brief review of the format of the assessments, the range achievement level descriptors (RALDs), and how the panelists were to make their judgments.

Following the general orientation, the facilitators reviewed the assessment's purpose, format (e.g., item types), and blueprint. After this introduction, the panelists were instructed to review sample items from the assessment from the NSCAS Item Type Samplers. The purpose of this review was to understand a student's experience interacting with the assessment itself. Panelists were then asked to review the RALDs, which describe the knowledge, skills, and ability (KSA) expectations for each achievement level that are tied to the grade-level content standards for an assessment. After the review, the panelists collaborated within their table groups to identify which KSAs they expected a student to perform if they were at the threshold of each achievement level. Each small group was assigned a specific domain to focus on when defining the threshold ALDs. Once all domains were covered, the facilitators reviewed the results with the whole panel and guided them through a discussion focused on refining the document until a final consensus was reached.

Next, the facilitators provided additional training on the Bookmark method. This training began with a conceptual review of how the panelists were to translate the expectations outlined in the threshold ALDs into progress within the OIB. The facilitators described the OIB and how it was created to consist of a set of items placed in order from least to most difficult using a response probability of two-thirds (0.67) and used data from the Spring 2023 (and previous) assessments to determine the item-level values.

The panelists were instructed how to access the OIBs through the online NWEA assessment system, and the facilitators reviewed how dichotomous and polytomous items were to be presented in the OIB, as well as how to access the scoring rubrics that were provided for reference.

Following the training, the panelists then had the opportunity to practice the Bookmark method by applying the description for a threshold *On-Track* student to a shorter practice OIB. Once the panelists had completed making their practice ratings, the facilitators led the entire panel in a discussion of the results, and the panelists were allotted time to ask any questions that might have come up during the practice.

After the panelists completed the Bookmark method practice and felt ready to complete operational judgments, they were asked to complete the *Readiness Evaluation* form, which asked them to indicate how ready they feel to proceed with the operational standard setting judgments.

Following the confirmation that all panelists understood the procedures and were prepared to make their operational judgments, panelists were instructed to begin reviewing the OIB for a lower grade level and making their Round 1 Bookmark judgments. The panelists were reminded that the process of making judgments is an individual activity but that they would be provided ample time to discuss items after all Round 1 judgments were completed.

After Round 1, the panelists were provided feedback that included summary statistics of the panel recommendations and a graphical depiction of the individual recommendations within the panel. The panelists were first instructed to discuss their reactions to the feedback within their small groups and then asked to share their small group's key discussion points with the whole panel. Throughout the discussion, the panelists were prompted to consider whether they were grounding their Round 1 judgments in how they expect a student should be able to perform or in the expectations defined in the threshold ALDs of how a threshold student is likely to perform. Throughout the discussion, the facilitator displayed specific items and asked panelists to discuss how they reached a judgment using the expectations defined in the threshold ALDs. After the Round 1 discussion, the facilitator reviewed instructions for making the Bookmark judgments and instructed the panelists to consider their initial judgments, the Round 1 results, and the Round 1 discussions when making their Round 2 judgments.

Following Round 2, the panel was provided with the same type of feedback from Round 1 and was provided impact data, or the percentage of students who would be classified into each achievement level using the median Round 2 recommendation. After reviewing the feedback, panelists once again discussed their reactions in their table groups and then in the whole group setting. Panelists discussed whether they thought the results presented were an accurate depiction of all Nebraska students.

Panelists were then given a review of how to provide Round 3 judgments, which were their final judgments for that grade level. Following the completion of Round 3, the results (the same feedback data that was provided in Round 2) were shared with the panelists for review.

7.1.3. Articulation

After the final round of standard setting, three representatives from each panel were invited to participate in a standards articulation process. During the articulation process, the panelists evaluated whether the cross-grade and cross-subject impact represented a reasonable and coherent set of results. The articulation process was anchored on two underlying principles:

- Achievement level expectations should be coherent across grades and subjects. This does not mean they need to match or follow a specific pattern but rather that they should be reasonable.
- The judgments of the standard setting panels should be honored, unless doing so would clearly violate the principle above.

The primary question for the panelists to consider was whether the magnitude and pattern of the impact data match the magnitude of the shifts and expectations from a content/standards perspective.

Immediately following the standard setting meeting, ACS presented the results to the Assessment and Accountability Advisory Committee (AAAC), who discussed the impact of the recommended cut scores and then made additional recommendations for final adjustments to improve coherence across the grades.

7.2. Science

In 2023, NWEA contracted with ACS Ventures, LLC (ACS) to conduct a standards validation for the NSCAS for grades 5 and 8 in science in order to review the cut scores that were established for new NSCAS science assessments in 2022.

7.2.1. Methodology

Given the design of the assessment and how students navigate each task, ACS designed a process that paralleled how the standards were set in 2022 to guide panels of Nebraska educators through the process of validating the two cut scores that the 2022 standard setting panels recommended to be used to distinguish the three achievement levels (i.e., *Developing*, *On Track*, *Advanced*) described within the RALDs:

1. The cut score that differentiates *Developing* performance from *On-Track* performance (i.e., threshold *On Track*)
2. The cut score that differentiates *On-Track* performance from *Advanced* performance (i.e., threshold *Advanced*)

Specifically, NWEA applied the cut scores (set in 2022) to the 2023 test forms and provided ACS with the draft cut and item-level difficulty. ACS then identified which items each threshold student would likely answer correctly (i.e., they would answer the easiest items correctly up to the cut score). This identification was designed to help the panelists understand how students would meet the cut scores through their item-level performance. The panelists were then asked to judge if these performance expectations were reasonable or should be adjusted.

At the end of the study, the panelists participated in a vertical review where the recommended cut scores for the two grades were collectively reviewed to ensure coherence with expected student performance. The performance of students at the high school level (on the science ACT) was also considered during this discussion.

7.2.2. Meeting Process

The first part of the meeting served as an introduction to the general standards validation process. It began with a large-group general session with a welcome and introductions from NDE. The ACS lead facilitator provided a high-level orientation and training on the standard setting process that occurred in 2022 and the methodology to be followed for the validation. The overview also included a brief review of the assessments, the RALDs and threshold ALDs, and how the panelists were to make their judgments. After the general orientation session, panelists began their work within the grade-level panels.

First, the facilitator reviewed the purpose of the assessment, the format (e.g., item types), and the blueprint guiding the assessment development. Afterward, the panelists had the opportunity to review the 2023 form of the assessment.

Panelists were then asked to review the RALDs, which describe the knowledge, skills, and ability (KSA) expectations for each achievement level that are tied to the grade-level content standards for an assessment. After the review, the panelists were able to review the threshold ALDs created during the 2022 standard setting.

Next, the facilitator provided additional training on the standard validation process. This began with a review of how the cut score expectations were translated into item-level performance.

Operational judgments began once all panelists indicated that they understood the procedures and were prepared to make their Round 1 judgments. Following that confirmation, panelists made their judgments for all items. After Round 1, feedback was provided that included a summary of the panelist recommendations, the difficulty of each item on the test form, the impact of the current cut scores, as well as the recommended changes to the cut scores. Panelists then had the opportunity to make a second (and final) round of judgments that indicated any recommended changes to the cut scores in consideration of the feedback they received. Following Round 2, panelists heard the results from their panel and completed an evaluation of the results.

After the final round of standard validation, the panelists participated in a vertical articulation meeting. During this meeting, panelists evaluated whether the cross-grade impact represented a reasonable set of expectations from grade 5 and grade 8.

Immediately following the standard validation meeting, ACS presented the results to the Assessment and Accountability Advisory Committee (AAAC) for review. ACS captured the feedback from this group for inclusion in this report.

7.3. Final Results

The recommended cut scores were presented to the Nebraska State Board of Education on August 4, 2023. Table 7.1 presents the final approved cut scores that were used for subsequent scoring (i.e., the cuts used starting from Spring 2023).

Table 7.1. Final Approved Cut Scores

Content Area	Grade	Cuts (Scale Scores)	
		<i>Developing/ On Track</i>	<i>On Track/ Advanced</i>
ELA	3	2443	2536
	4	2493	2567
	5	2504	2583
	6	2518	2594
	7	2527	2609
	8	2524	2624
Math	3	1176	1297
	4	1208	1331
	5	1207	1320
	6	1228	1321
	7	1212	1314
	8	1231	1319
Science	5	3100	3150
	8	3100	3150

Section 8: Test Results

All students who took the online, paper-pencil, and Spanish forms of the Spring 2023 NSCAS Growth assessments were included in the test results. For results based on demographics and accommodations, all participants (i.e., students who attempted at least one item) were included. For all other results in this section, students who attempted at least 10 operational items on the online and paper-pencil forms were included. Results presented in this section are not from the state student file that NDE received and may, therefore, differ slightly from the official state summary report due to ongoing resolution of test materials and slight differences in the application of exclusion rules.

8.1. Demographics and Accommodations

Table 8.1–Table 8.6 present the number of tested students by demographics for each grade and content area, including gender, ethnicity, free and reduced lunch (FRL) status, limited English proficiency (LEP) status, special education (SPED) status, use of universal features (i.e., answer eliminator, highlighter, notepad, and zoom), and use of accommodations (i.e., text-to-speech [TTS], paper-pencil form, Spanish online or paper-pencil form, Braille, and large print). Starting in 2018, both current and former English language learner (ELL) students are considered to have LEP status, resulting in more LEP students compared with previous years.

As shown in these tables, approximately 23,000 students took the assessment in each grade and content area. Of those students across grades, half are males, half are females, two thirds are white, and about one-fifth are Hispanic. Among the students across grades, about 44–47% are eligible for FRL, 9–17% have LEP status, and 14–17% belong to at least one SPED category.

Table 8.1. Number of Students Tested by Demographics and Accommodations—Grade 3

Demographic Subgroup		ELA		Mathematics	
		N	%	N	%
Total N-Count		23,282	100	23,285	100
Gender	Female	11,431	49.1	11,433	49.1
	Male	11,851	50.9	11,852	50.9
Ethnicity	AI/AN	274	1.18	274	1.18
	Asian	762	3.27	761	3.27
	Black or African American	1,479	6.35	1,479	6.35
	Hispanic	4,961	21.31	4,957	21.29
	NH/PI	34	0.15	34	0.15
	White	14,631	62.85	14,638	62.88
	Two or More Races	1,137	4.88	1,138	4.89
FRL	Yes	11,071	47.56	11,063	47.52
	No	12,207	52.44	12,218	52.48
LEP	Yes	4,126	17.72	4,120	17.7
	No	19,153	82.28	19,162	82.3
SPED	Yes	4,382	18.82	4,381	18.81

Demographic Subgroup		ELA		Mathematics	
		N	%	N	%
	No	18,900	81.18	18,904	81.19
Universal Features & Accommodations	Text to Speech	4,610	19.8	4,845	20.81
	Basic Calculator	—	—	925	3.97
	Read Aloud	171	0.73	154	0.66
	One-on-One Setting	1,244	5.34	1,238	5.32
	Bilingual Dictionary/Word List	20	0.09	75	0.32
	Language Translation	—	—	84	0.36
	Mathematical Supports	—	—	955	4.1
	Assistive Technology	30	0.13	29	0.12
	Specialized Presentation	8	0.03	8	0.03
	Scribe	45	0.19	46	0.2
	Paper-Pencil (PP)	5	0.02	—	—
	Spanish Online	20	0.09	87	0.37
	Spanish Paper-Pencil (PP)	—	—	—	—
	Braille ^a	0	—	—	—
	Large Print ^a	1	—	—	—

Note. AI/AN = American Indian or Alaskan Native; NH/PI = Native Hawaiian or Other Pacific Islander; FRL = free and reduced lunch; LEP = limited English proficient; SPED = special education

^a Braille and large print counts are based on students who actually tested and were not included in the total n-count.

Table 8.2. Number of Students Tested by Demographics and Accommodations—Grade 4

Demographic Subgroup		ELA		Mathematics	
		N	%	N	%
Total N-Count		22,957	100	22,947	100
Gender	Female	11,175	48.68	11,169	48.67
	Male	11,782	51.32	11,778	51.33
Ethnicity	AI/AN	273	1.19	275	1.2
	Asian	776	3.38	776	3.38
	Black or African American	1,476	6.43	1,475	6.43
	Hispanic	4,868	21.21	4,863	21.19
	NH/PI	38	0.17	38	0.17
	White	14,388	62.68	14,386	62.7
	Two or More Races	1,134	4.94	1,132	4.93
FRL	Yes	10,865	47.34	10,857	47.32
	No	12,088	52.66	12,088	52.68
LEP	Yes	3,938	17.16	3,934	17.15
	No	19,017	82.84	19,011	82.85
SPED	Yes	4,086	17.8	4,087	17.81

Demographic Subgroup		ELA		Mathematics	
		N	%	N	%
	No	18,871	82.2	18,860	82.19
Universal Features & Accommodations	Text to Speech	4,363	19.01	4,575	19.94
	Basic Calculator	—	—	1,029	4.48
	Read Aloud	183	0.8	175	0.76
	One-on-One Setting	1,255	5.47	1,256	5.47
	Bilingual Dictionary/Word List	24	0.1	56	0.24
	Language Translation	—	—	66	0.29
	Mathematical Supports	—	—	1,123	4.89
	Assistive Technology	19	0.08	19	0.08
	Specialized Presentation	9	0.04	9	0.04
	Scribe	58	0.25	54	0.24
	Paper-Pencil (PP)	7	0.03	6	0.03
	Spanish Online	37	0.16	104	0.45
	Spanish Paper-Pencil (PP)	—	—	—	—
	Braille ^a	0	—	0	—
	Large Print ^a	4	—	3	—

Note. AI/AN = American Indian or Alaskan Native; NH/PI = Native Hawaiian or Other Pacific Islander; FRL = free and reduced lunch; LEP = limited English proficient; SPED = special education

^a Braille and large print counts are based on students who actually tested and were not included in the total n-count.

Table 8.3. Number of Students Tested by Demographics and Accommodations—Grade 5

Demographic Subgroup		ELA		Mathematics		Science	
		N	%	N	%	N	%
Total N-Count		23,031	100	23,026	100	22,998	100
Gender	Female	11,163	48.47	11,162	48.48	11,151	48.49
	Male	11,868	51.53	11,864	51.52	11,847	51.51
Ethnicity	AI/AN	290	1.26	291	1.26	286	1.24
	Asian	746	3.24	745	3.24	746	3.24
	Black or African American	1,444	6.27	1,443	6.27	1,444	6.28
	Hispanic	4,715	20.47	4,712	20.46	4,712	20.49
	NH/PI	48	0.21	48	0.21	47	0.2
	White	14,645	63.59	14,645	63.6	14,621	63.58
	Two or More Races	1,141	4.95	1,141	4.96	1,139	4.95
FRL	Yes	10,598	46.02	10,594	46.01	10,575	46
	No	12,429	53.98	12,429	53.99	12,416	54
LEP	Yes	3,393	14.73	3,391	14.73	3,385	14.72
	No	19,636	85.27	19,634	85.27	19,611	85.28
SPED	Yes	3,899	16.93	3,899	16.93	3,895	16.94

Demographic Subgroup		ELA		Mathematics		Science	
		N	%	N	%	N	%
	No	19,132	83.07	19,127	83.07	19,103	83.06
Universal Features & Accommodations	Text to Speech	3,996	17.35	4,125	17.91	4,084	17.76
	Basic Calculator	–	–	1,236	5.37	–	–
	Read Aloud	173	0.75	166	0.72	171	0.74
	One-on-One Setting	1,204	5.23	1,198	5.2	1,189	5.17
	Bilingual Dictionary/Word List	41	0.18	56	0.24	58	0.25
	Language Translation	–	–	56	0.24	62	0.27
	Mathematical Supports	–	–	1,249	5.42	–	–
	Assistive Technology	26	0.11	25	0.11	21	0.09
	Specialized Presentation	15	0.07	19	0.08	12	0.05
	Scribe	39	0.17	35	0.15	32	0.14
	Paper-Pencil (PP)	3	0.01	3	0.01	2	0.01
	Spanish Online	51	0.22	104	0.45	105	0.46
	Spanish Paper-Pencil (PP)	–	–	1	0	1	0
	Braille ^a	2	–	2	–	2	–
	Large Print ^a	1	–	1	–	1	–

Note. AI/AN = American Indian or Alaskan Native; NH/PI = Native Hawaiian or Other Pacific Islander; FRL = free and reduced lunch; LEP = limited English proficient; SPED = special education

^a Braille and large print counts are based on students who actually tested and were not included in the total n-count.

Table 8.4. Number of Students Tested by Demographics and Accommodations—Grade 6

Demographic Subgroup		ELA		Mathematics	
		N	%	N	%
Total N-Count		22,889	100	22,870	100
Gender	Female	11,140	48.67	11,138	48.7
	Male	11,749	51.33	11,732	51.3
Ethnicity	AI/AN	255	1.11	254	1.11
	Asian	708	3.09	707	3.09
	Black or African American	1,425	6.23	1,418	6.2
	Hispanic	4,721	20.63	4,718	20.63
	NH/PI	41	0.18	41	0.18
	White	14,647	64.01	14,642	64.03
	Two or More Races	1,087	4.75	1,087	4.75
FRL	Yes	10,192	44.54	10,182	44.53

Demographic Subgroup		ELA		Mathematics	
		N	%	N	%
	No	12,691	55.46	12,683	55.47
LEP	Yes	2,852	12.46	2,851	12.47
	No	20,033	87.54	20,015	87.53
SPED	Yes	3,584	15.66	3,579	15.65
	No	19,305	84.34	19,291	84.35
Universal Features & Accommodations	Text to Speech	3,577	15.63	3,627	15.86
	Basic Calculator	–	–	1,691	7.39
	Read Aloud	115	0.5	115	0.5
	One-on-One Setting	901	3.94	917	4.01
	Bilingual Dictionary/Word List	17	0.07	60	0.26
	Language Translation	–	–	54	0.24
	Mathematical Supports	–	–	1,153	5.04
	Assistive Technology	23	0.1	23	0.1
	Specialized Presentation	9	0.04	9	0.04
	Scribe	35	0.15	37	0.16
	Paper-Pencil (PP)	6	0.03	6	0.03
	Spanish Online	33	0.14	86	0.38
	Spanish Paper-Pencil (PP)	–	–	–	–
	Braille ^a	3	–	3	–
	Large Print ^a	3	–	3	–

Note. AI/AN = American Indian or Alaskan Native; NH/PI = Native Hawaiian or Other Pacific Islander; FRL = free and reduced lunch; LEP = limited English proficient; SPED = special education

^a Braille and large print counts are based on students who actually tested and were not included in the total n-count.

Table 8.5. Number of Students Tested by Demographics and Accommodations—Grade 7

Demographic Subgroup		ELA		Mathematics	
		N	%	N	%
Total N-Count		23,493	100	23,463	100
Gender	Female	11,454	48.75	11,436	48.74
	Male	12,039	51.25	12,027	51.26
Ethnicity	AI/AN	286	1.22	287	1.22
	Asian	707	3.01	705	3.01
	Black or African American	1,560	6.64	1,563	6.66
	Hispanic	4,920	20.95	4,915	20.95
	NH/PI	43	0.18	42	0.18
	White	14,873	63.33	14,853	63.32
	Two or More Races	1,095	4.66	1,091	4.65

Demographic Subgroup		ELA		Mathematics	
		N	%	N	%
FRL	Yes	10,472	44.6	10,455	44.57
	No	13,010	55.4	13,000	55.43
LEP	Yes	2,489	10.6	2,483	10.58
	No	20,998	89.4	20,975	89.42
SPED	Yes	3,379	14.38	3,372	14.37
	No	20,114	85.62	20,091	85.63
Universal Features & Accommodations	Text to Speech	3,085	13.13	3,108	13.25
	Scientific Calculator	—	—	1,499	6.39
	Read Aloud	113	0.48	107	0.46
	One-on-One Setting	938	3.99	940	4.01
	Bilingual Dictionary/Word List	27	0.11	54	0.23
	Language Translation	—	—	66	0.28
	Mathematical Supports	—	—	946	4.03
	Assistive Technology	17	0.07	18	0.08
	Specialized Presentation	6	0.03	6	0.03
	Scribe	15	0.06	14	0.06
	Paper-Pencil (PP)	5	0.02	4	0.02
	Spanish Online	58	0.25	110	0.47
	Spanish Paper-Pencil (PP)	—	—	—	—
	Braille ^a	2	—	2	—
	Large Print ^a	1	—	1	—

Note. AI/AN = American Indian or Alaskan Native; NH/PI = Native Hawaiian or Other Pacific Islander; FRL = free and reduced lunch; LEP = limited English proficient; SPED = special education

^a Braille and large print counts are based on students who actually tested and were not included in the total n-count.

Table 8.6. Number of Students Tested by Demographics and Accommodations—Grade 8

Demographic Subgroup		ELA		Mathematics		Science	
		N	%	N	%	N	%
Total N-Count		23,956	100	23,919	100	23,939	100
Gender	Female	11,645	48.61	11,639	48.66	11,636	48.61
	Male	12,311	51.39	12,280	51.34	12,303	51.39
Ethnicity	AI/AN	303	1.27	301	1.26	301	1.26
	Asian	664	2.77	662	2.77	663	2.77
	Black or African American	1,546	6.45	1,541	6.44	1,546	6.46
	Hispanic	5,213	21.76	5,212	21.79	5,210	21.77
	NH/PI	35	0.15	33	0.14	34	0.14
	White	15,141	63.21	15,115	63.21	15,130	63.22

Demographic Subgroup		ELA		Mathematics		Science	
		N	%	N	%	N	%
	Two or More Races	1,050	4.38	1,050	4.39	1,048	4.38
FRL	Yes	10,431	43.55	10,411	43.54	10,413	43.51
	No	13,520	56.45	13,502	56.46	13,518	56.49
LEP	Yes	2,261	9.44	2,257	9.44	2,262	9.45
	No	21,691	90.56	21,656	90.56	21,672	90.55
SPED	Yes	3,287	13.72	3,281	13.72	3,286	13.73
	No	20,669	86.28	20,638	86.28	20,653	86.27
Universal Features & Accommodations	Text to Speech	3,023	12.62	3,064	12.81	3,066	12.81
	Scientific Calculator	–	–	1,640	6.86	–	–
	Read Aloud	111	0.46	104	0.43	106	0.44
	One-on-One Setting	829	3.46	818	3.42	814	3.4
	Bilingual Dictionary/Word List	29	0.12	79	0.33	68	0.28
	Language Translation	–	–	66	0.28	67	0.28
	Mathematical Supports	–	–	895	3.74	–	–
	Assistive Technology	26	0.11	22	0.09	15	0.06
	Specialized Presentation	7	0.03	7	0.03	8	0.03
	Scribe	10	0.04	10	0.04	8	0.03
	Paper-Pencil (PP)	16	0.07	17	0.07	17	0.07
	Spanish Online	59	0.25	121	0.51	126	0.53
	Spanish Paper-Pencil (PP)	–	–	–	–	–	–
	Braille ^a	2	–	2	–	2	–
	Large Print ^a	4	–	5	–	4	–

Note. AI/AN = American Indian or Alaskan Native; NH/PI = Native Hawaiian or Other Pacific Islander; FRL = free and reduced lunch; LEP = limited English proficient; SPED = special education

^a Braille and large print counts are based on students who actually tested and were not included in the total n-count.

8.2. Administration Mode (Online vs. Paper-Pencil)

The 2023 NSCAS assessments were administered online to the extent practical, and a very small number of students took the paper-pencil test. As shown in Table 8.7, less than 1% of students took the assessment in the paper-based version across all grades and content areas.

Table 8.7. Number of Students Tested by Administration Mode

Content Area	Grade	Total #Students	Online N	Paper-Pencil	
				N	%
ELA	3	23,260	23,255	5	0.0
	4	22,918	22,911	7	0.0
	5	22,977	22,974	3	0.0
	6	22,851	22,845	6	0.0
	7	23,430	23,425	5	0.0
	8	23,886	23,870	16	0.1

Content Area	Grade	Total #Students	Online N	Paper-Pencil	
				N	%
Mathematics	3	23,197	23,197	0	0.0
	4	22,842	22,836	6	0.0
	5	22,917	22,914	3	0.0
	6	22,774	22,768	6	0.0
	7	23,348	23,344	4	0.0
	8	23,787	23,771	16	0.1
Science	5	22,888	22,886	2	0.0
	8	23,807	23,790	17	0.1

8.3. Testing Time

Table 8.8 through Table 8.10 present the numbers of minutes students spent taking the Spring 2023 NSCAS ELA, mathematics, and science assessments, respectively. Specifically, the tables present the numbers and percentages of students who completed the tests in various time ranges. As shown in the tables, most students finished the tests within 120 minutes, and the percentage of students who took more than 180 minutes is less than 2%.

Table 8.8. Testing Time in Minutes—ELA

Time (in minutes)	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8	
	N	%	N	%	N	%	N	%	N	%	N	%
<10	52	0.2	30	0.1	37	0.2	51	0.2	65	0.3	80	0.3
10—<20	380	1.6	201	0.9	193	0.8	242	1.1	247	1.1	295	1.2
20—<30	926	4.0	594	2.6	552	2.4	534	2.3	726	3.1	815	3.4
30—<40	1,724	7.4	1,447	6.3	1,310	5.7	1,257	5.5	1,512	6.5	1,889	7.9
40—<50	2,647	11.4	2,411	10.5	2,386	10.4	2,342	10.3	2,786	11.9	3,264	13.7
50—<60	3,215	13.8	3,128	13.7	3,252	14.2	3,283	14.4	3,670	15.7	4,047	17.0
60—<70	3,124	13.4	3,208	14.0	3,440	15.0	3,659	16.0	3,801	16.2	4,044	16.9
70—<80	2,835	12.2	2,979	13.0	3,075	13.4	3,263	14.3	3,281	14.0	3,251	13.6
80—<90	2,296	9.9	2,442	10.7	2,531	11.0	2,568	11.2	2,427	10.4	2,251	9.4
90—<100	1,729	7.4	1,937	8.5	1,977	8.6	1,859	8.1	1,639	7.0	1,488	6.2
100—<110	1,266	5.4	1,362	5.9	1,261	5.5	1,237	5.4	1,174	5.0	872	3.7
110—<120	893	3.8	980	4.3	971	4.2	887	3.9	745	3.2	563	2.4
120—<130	632	2.7	674	2.9	598	2.6	548	2.4	446	1.9	343	1.4
130—<140	427	1.8	480	2.1	421	1.8	341	1.5	323	1.4	202	0.8
140—<150	301	1.3	289	1.3	278	1.2	256	1.1	205	0.9	159	0.7
150—<160	224	1.0	219	1.0	217	0.9	149	0.7	116	0.5	101	0.4
160—<170	155	0.7	158	0.7	130	0.6	98	0.4	74	0.3	76	0.3
170—<180	112	0.5	116	0.5	103	0.4	93	0.4	57	0.2	35	0.1
≥180	317	1.4	256	1.1	242	1.1	178	0.8	131	0.6	95	0.4
Total	23,255	100.0	22,911	100.0	22,974	100.0	22,845	100.0	23,425	100.0	23,870	100.0

Table 8.9. Testing Time in Minutes—Mathematics

Time (in minutes)	Grade 3		Grade 4		Grade 5		Grade 6		Grade 7		Grade 8	
	N	%	N	%	N	%	N	%	N	%	N	%
<10	11	0.0	8	0.0	13	0.1	27	0.1	43	0.2	83	0.3
10—<20	238	1.0	128	0.6	134	0.6	208	0.9	230	1.0	319	1.3
20—<30	999	4.3	538	2.4	503	2.2	652	2.9	609	2.6	855	3.6
30—<40	2,600	11.2	1,652	7.2	1,521	6.6	1,363	6.0	1,459	6.3	1,794	7.5
40—<50	4,000	17.2	2,953	12.9	2,806	12.2	2,251	9.9	2,301	9.9	2,836	11.9
50—<60	4,114	17.7	3,499	15.3	3,584	15.6	2,882	12.7	3,059	13.1	3,591	15.1
60—<70	3,451	14.9	3,372	14.8	3,584	15.6	2,994	13.2	3,383	14.5	3,613	15.2
70—<80	2,380	10.3	2,846	12.5	3,016	13.2	2,784	12.2	3,074	13.2	3,030	12.7
80—<90	1,719	7.4	2,179	9.5	2,293	10.0	2,437	10.7	2,648	11.3	2,318	9.8
90—<100	1,164	5.0	1,659	7.3	1,640	7.2	1,870	8.2	1,970	8.4	1,747	7.3
100—<110	809	3.5	1,136	5.0	1,149	5.0	1,365	6.0	1,382	5.9	1,193	5.0
110—<120	462	2.0	836	3.7	783	3.4	1,105	4.9	963	4.1	804	3.4
120—<130	352	1.5	604	2.6	540	2.4	771	3.4	672	2.9	491	2.1
130—<140	234	1.0	393	1.7	387	1.7	606	2.7	484	2.1	362	1.5
140—<150	170	0.7	269	1.2	269	1.2	399	1.8	295	1.3	241	1.0
150—<160	116	0.5	211	0.9	225	1.0	294	1.3	224	1.0	157	0.7
160—<170	111	0.5	173	0.8	140	0.6	213	0.9	147	0.6	105	0.4
170—<180	66	0.3	106	0.5	87	0.4	139	0.6	105	0.4	53	0.2
≥180	201	0.9	274	1.2	240	1.0	408	1.8	296	1.3	179	0.8
Total	23,197	100.0	22,836	100.0	22,914	100.0	22,768	100.0	23,344	100.0	23,771	100.0

Table 8.10. Testing Time in Minutes—Science

Time (in minutes)	Grade 5		Grade 8	
	N	%	N	%
<10	38	0.2	115	0.5
10–<20	268	1.2	592	2.5
20–<30	1,201	5.2	2,119	8.9
30–<40	3,394	14.8	4,965	20.9
40–<50	4,711	20.6	6,138	25.8
50–<60	4,553	19.9	4,374	18.4
60–<70	3,335	14.6	2,595	10.9
70–<80	2,124	9.3	1,257	5.3
80–<90	1,325	5.8	668	2.8
90–<100	753	3.3	407	1.7
100–<110	472	2.1	227	1.0
110–<120	266	1.2	120	0.5
120–<130	173	0.8	90	0.4
130–<140	90	0.4	33	0.1
140–<150	65	0.3	31	0.1
150–<160	47	0.2	23	0.1
160–<170	26	0.1	8	0.0
170–<180	15	0.1	4	0.0
≥180	30	0.1	24	0.1
Total	22,886	100.0	23,790	100.0

8.4. Achievement Level Distributions

Table 8.11 presents the achievement level distributions for the Spring 2023 NSCAS assessments. Appendix D: Achievement Level Distributions & Scale Score Descriptive Statistics by Demographics provides the achievement level distributions by demographic group. For ELA, 37–46% of students are at *Developing*, and 54–66% of students are at *On Track* or *Advanced*. For mathematics, 34–42% of students are at *Developing*, and 58–66% of students are at *On Track* or *Advanced*. For science, 23–35% of students are at *Developing*, and 65–77% are at *On Track* or *Advanced*.

Table 8.11. Achievement Level Distributions

Content Area	Grade	Total N-Count	Level 3		Level 2		Level 1		Level 2 + Level 1	
			N-Count	%	N-Count	%	N-Count	%	N-Count	%
ELA	3	23,260	8,766	37.7	9,428	40.5	5,066	21.8	14,494	62.3
	4	22,918	10,306	45.0	7,573	33.0	5,039	22.0	12,612	55.0
	5	22,977	9,917	43.2	8,312	36.2	4,748	20.7	13,060	56.8
	6	22,851	10,204	44.7	8,622	37.7	4,025	17.6	12,647	55.3
	7	23,430	10,723	45.8	9,069	38.7	3,638	15.5	12,707	54.2
	8	23,886	8,720	36.5	11,424	47.8	3,742	15.7	15,166	63.5
Math	3	23,197	9,695	41.8	10,452	45.1	3,050	13.1	13,502	58.2
	4	22,842	9,585	42.0	10,500	46.0	2,757	12.1	13,257	58.0
	5	22,917	7,939	34.6	10,659	46.5	4,319	18.8	14,978	65.4
	6	22,774	9,641	42.3	9,114	40.0	4,019	17.6	13,133	57.7
	7	23,348	7,932	34.0	10,736	46.0	4,680	20.0	15,416	66.0
	8	23,787	9,064	38.1	9,214	38.7	5,509	23.2	14,723	61.9
Science	5	22,888	5,248	22.9	13,985	61.1	3,655	16.0	17,640	77.1
	8	23,807	8,305	34.9	13,480	56.6	2,022	8.5	15,502	65.1

Note. Achievement levels: Level 3 = *Developing*; Level 2 = *On Track*; Level 1 = *Advanced*

8.5. Descriptive Statistics of Scale Scores

Table 8.12 presents the descriptive statistics for the scale scores, including the mean, standard deviation (SD), and scores at the 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentiles. Appendix D: Achievement Level Distributions & Scale Score Descriptive Statistics by Demographics also presents the descriptive statistics by demographic group. The mean scale score increases with the grade levels for ELA and mathematics, as expected.

Table 8.12. Scale Score Descriptive Statistics

Content Area	Gr.	N-Count	LOSS	HOSS	Min	Max	Mean	SD	Percentiles						
									P5	P10	P25	P50	P75	P90	P95
ELA	3	23,260	2220	2840	2222	2840	2463.42	90.77	2299	2334	2406	2471	2527	2575	2601
	4	22,918	2250	2850	2252	2844	2493.26	92.39	2320	2362	2436	2504	2559	2602	2631
	5	22,977	2280	2860	2282	2851	2510.98	87.99	2356	2392	2452	2519	2570	2620	2646
	6	22,851	2290	2870	2292	2780	2518.30	81.57	2367	2401	2465	2529	2575	2616	2638
	7	23,430	2300	2880	2302	2809	2527.56	81.88	2376	2414	2476	2536	2586	2626	2650
	8	23,886	2310	2890	2312	2849	2544.79	80.09	2397	2434	2496	2551	2601	2642	2667
Math	3	23,197	1000	1470	1002	1470	1193.73	88.26	1042	1078	1135	1193	1249	1314	1344
	4	22,842	1010	1500	1012	1500	1224.02	87.18	1078	1106	1164	1224	1283	1341	1368
	5	22,917	1020	1510	1022	1510	1242.12	83.43	1107	1139	1183	1239	1301	1355	1381
	6	22,774	1030	1530	1032	1530	1242.68	85.48	1091	1129	1190	1243	1298	1352	1387
	7	23,348	1040	1540	1042	1540	1246.44	83.63	1112	1140	1190	1242	1299	1355	1395
	8	23,787	1050	1550	1052	1550	1254.57	88.19	1104	1133	1198	1255	1313	1368	1402
Science	5	22,888	3000	3250	3008	3226	3119.63	27.89	3078	3082	3100	3120	3140	3156	3163
	8	23,807	3000	3250	3002	3233	3111.58	30.21	3060	3070	3090	3114	3132	3149	3164

8.6. Reporting Category Correlations

For each grade and content area, Pearson's correlation coefficients between reporting category scores were calculated to provide information on score dimensionality, which is part of validity evidence based on the tests' internal structure. Disattenuated correlations provide an estimate of the relationships between reporting categories if there is no measurement error. Table 8.13 provides the reporting category correlations, and Table 8.14 presents the disattenuated correlations.

The correlations between reporting categories *within* the content areas are positive and moderate in value (i.e., higher than 0.60), while the correlations between reporting categories *across* the content areas are positive and low to moderate in value (i.e., higher than 0.50). In general, the within-content-area reporting category correlations are higher than the across-content-area reporting category correlations.

The disattenuated correlations are higher than the correlations, which is expected given that none of the reporting categories has perfect reliabilities (see Table 9.1–Table 9.3). The disattenuated correlations between reporting categories *within* the content areas are positive and high in value (i.e., higher than 0.80), while the disattenuated correlations between reporting categories *across* the content areas are positive and moderate in value (i.e., higher than 0.60). These ranges are similar to those from last year. The high disattenuated correlations within the content suggest that reporting categories might be measuring essentially the same construct, which is one piece of evidence based on internal structure. In other words, the internal structure of the assessments is consistent with the structure of the content standards.

Table 8.13. Reporting Category Correlations

Grade	Reporting Category	Reporting Category							
		E1	E2	E3	E4	M1	M2	M3	M4
3	RP (E1)	1.00							
	RI (E2)	0.69	1.00						
	Vocabulary (E3)	0.61	0.64	1.00					
	Writing (E4)	0.62	0.63	0.56	1.00				
	Number (M1)	0.60	0.62	0.57	0.57	1.00			
	Algebra (M2)	0.58	0.60	0.54	0.55	0.77	1.00		
	Geometry (M3)	0.59	0.62	0.57	0.57	0.79	0.74	1.00	
	Data (M4)	0.60	0.62	0.57	0.57	0.77	0.73	0.74	1.00
4	RP (E1)	1.00							
	RI (E2)	0.67	1.00						
	Vocabulary (E3)	0.61	0.64	1.00					
	Writing (E4)	0.63	0.64	0.59	1.00				
	Number (M1)	0.56	0.59	0.53	0.56	1.00			
	Algebra (M2)	0.59	0.62	0.57	0.59	0.78	1.00		
	Geometry (M3)	0.54	0.58	0.53	0.55	0.74	0.72	1.00	
	Data (M4)	0.54	0.57	0.52	0.54	0.72	0.72	0.68	1.00
5	RP (E1)	1.00							
	RI (E2)	0.66	1.00						

Grade	Reporting Category	Reporting Category							
		E1	E2	E3	E4	M1	M2	M3	M4
	Vocabulary (E3)	0.60	0.62	1.00					
	Writing (E4)	0.63	0.64	0.57	1.00				
	Number (M1)	0.56	0.58	0.54	0.56	1.00			
	Algebra (M2)	0.56	0.57	0.53	0.56	0.77	1.00		
	Geometry (M3)	0.54	0.56	0.52	0.55	0.74	0.70	1.00	
	Data (M4)	0.58	0.60	0.55	0.59	0.72	0.69	0.67	1.00
6	RP (E1)	1.00							
	RI (E2)	0.64	1.00						
	Vocabulary (E3)	0.56	0.58	1.00					
	Writing (E4)	0.58	0.61	0.52	1.00				
	Number (M1)	0.56	0.59	0.52	0.54	1.00			
	Algebra (M2)	0.59	0.63	0.55	0.58	0.81	1.00		
	Geometry (M3)	0.52	0.56	0.48	0.51	0.72	0.74	1.00	
	Data (M4)	0.54	0.57	0.50	0.53	0.72	0.75	0.68	1.00
7	RP (E1)	1.00							
	RI (E2)	0.64	1.00						
	Vocabulary (E3)	0.58	0.58	1.00					
	Writing (E4)	0.60	0.62	0.55	1.00				
	Number (M1)	0.52	0.56	0.49	0.54	1.00			
	Algebra (M2)	0.58	0.61	0.55	0.59	0.76	1.00		
	Geometry (M3)	0.53	0.57	0.50	0.54	0.71	0.76	1.00	
	Data (M4)	0.55	0.59	0.53	0.57	0.71	0.76	0.72	1.00
8	RP (E1)	1.00							
	RI (E2)	0.63	1.00						
	Vocabulary (E3)	0.56	0.57	1.00					
	Writing (E4)	0.58	0.60	0.53	1.00				
	Number (M1)	0.52	0.54	0.49	0.53	1.00			
	Algebra (M2)	0.56	0.58	0.52	0.57	0.77	1.00		
	Geometry (M3)	0.55	0.56	0.51	0.55	0.76	0.78	1.00	
	Data (M4)	0.54	0.56	0.50	0.55	0.69	0.73	0.72	1.00

Note. E1 = Reading Prose and Poetry; E2 = Reading Informational Text; E4 = Writing and Foundations of Writing

Table 8.14. Reporting Category Disattenuated Correlations

Grade	Reporting Category	Reporting Category							
		E1	E2	E3	E4	M1	M2	M3	M4
3	RP (E1)	1.00							
	RI (E2)	0.93	1.00						
	Vocabulary (E3)	0.99	1.00	1.00					
	Writing (E4)	0.91	0.91	0.96	1.00				
	Number (M1)	0.76	0.78	0.84	0.77	1.00			
	Algebra (M2)	0.78	0.80	0.86	0.79	0.96	1.00		
	Geometry (M3)	0.78	0.80	0.87	0.80	0.95	0.96	1.00	

Grade	Reporting Category	Reporting Category							
		E1	E2	E3	E4	M1	M2	M3	M4
	Data (M4)	0.81	0.82	0.88	0.81	0.95	0.96	0.95	1.00
4	RP (E1)	1.00							
	RI (E2)	0.93	1.00						
	Vocabulary (E3)	1.00	1.00	1.00					
	Writing (E4)	0.91	0.91	1.00	1.00				
	Number (M1)	0.71	0.74	0.88	0.74	1.00			
	Algebra (M2)	0.78	0.81	0.98	0.81	0.94	1.00		
	Geometry (M3)	0.72	0.76	0.92	0.75	0.90	0.90	1.00	
	Data (M4)	0.74	0.76	0.91	0.76	0.90	0.92	0.88	1.00
5	RP (E1)	1.00							
	RI (E2)	0.93	1.00						
	Vocabulary (E3)	1.00	1.00	1.00					
	Writing (E4)	0.94	0.93	1.00	1.00				
	Number (M1)	0.73	0.73	0.83	0.76	1.00			
	Algebra (M2)	0.76	0.76	0.86	0.78	0.94	1.00		
	Geometry (M3)	0.73	0.74	0.85	0.77	0.91	0.90	1.00	
	Data (M4)	0.83	0.83	0.93	0.86	0.91	0.92	0.90	1.00
6	RP (E1)	1.00							
	RI (E2)	0.91	1.00						
	Vocabulary (E3)	1.00	1.00	1.00					
	Writing (E4)	0.89	0.90	0.99	1.00				
	Number (M1)	0.74	0.77	0.85	0.76	1.00			
	Algebra (M2)	0.77	0.79	0.89	0.80	0.97	1.00		
	Geometry (M3)	0.74	0.77	0.86	0.77	0.95	0.95	1.00	
	Data (M4)	0.75	0.77	0.86	0.77	0.92	0.93	0.93	1.00
7	RP (E1)	1.00							
	RI (E2)	0.93	1.00						
	Vocabulary (E3)	1.00	0.97	1.00					
	Writing (E4)	0.93	0.92	1.00	1.00				
	Number (M1)	0.73	0.75	0.81	0.78	1.00			
	Algebra (M2)	0.78	0.79	0.85	0.82	0.95	1.00		
	Geometry (M3)	0.75	0.77	0.83	0.79	0.94	0.95	1.00	
	Data (M4)	0.77	0.79	0.85	0.81	0.91	0.94	0.94	1.00
8	RP (E1)	1.00							
	RI (E2)	0.91	1.00						
	Vocabulary (E3)	1.00	1.00	1.00					
	Writing (E4)	0.89	0.90	1.00	1.00				
	Number (M1)	0.71	0.72	0.85	0.76	1.00			
	Algebra (M2)	0.75	0.76	0.89	0.79	0.95	1.00		
	Geometry (M3)	0.72	0.73	0.87	0.77	0.93	0.94	1.00	
	Data (M4)	0.76	0.78	0.90	0.81	0.91	0.95	0.92	1.00

Note. E1 = Reading Prose and Poetry; E2 = Reading Informational Text; E4 = Writing and Foundations of Writing

8.7. Correlations with MAP Growth

Table 8.15 presents the correlation coefficients between MAP Growth and NSCAS scores for students who took both tests in Spring 2023. As shown in the table, the correlation coefficients are higher than 0.80 for both ELA and mathematics. In general, these high correlations indicate that the relationship between MAP Growth and NSCAS test scores is strong, which can be considered validity evidence based on other variables.

Table 8.15. Correlation and Descriptive Statistics of NSCAS and MAP Growth Scores

Grade	N	r	NSCAS				MAP Growth			
			Mean	SD	Min.	Max.	Mean	SD	Min.	Max.
ELA										
3	6,780	0.84	2482	87.16	2224	2840	201	14.78	145	245
4	6,976	0.84	2510	87.49	2256	2844	209	14.61	142	252
5	6,637	0.83	2529	83.70	2283	2851	215	13.97	150	259
6	6,331	0.83	2535	77.44	2293	2755	218	14.20	156	271
7	5,077	0.82	2545	75.09	2304	2783	221	14.08	157	263
8	5,050	0.82	2560	75.18	2315	2795	224	14.59	161	269
Mathematics										
3	6,680	0.87	1214	84.05	1003	1470	206	13.13	130	257
4	6,714	0.89	1244	84.23	1013	1500	216	15.21	144	285
5	6,546	0.89	1261	81.36	1026	1510	225	17.23	145	290
6	6,385	0.89	1263	79.41	1034	1530	228	16.06	140	284
7	4,975	0.89	1268	81.19	1047	1540	232	17.53	152	303
8	5,071	0.87	1279	84.93	1057	1550	237	19.43	141	310

Note. r = correlation; SD = standard deviation; Min. = minimum; Max. = maximum

Section 9: Reliability

The *Standards for Educational and Psychological Testing* refers to reliability as the “consistency of scores across replications of a testing procedure” (AERA et al., 2014, p. 33). The level of reliability/precision of scores has implications for validity. In other words, scores must be consistent and precise enough to be useful for their intended purposes. If scores are to be meaningful, tests should produce stable scores if the same group of students were to take the same test repeatedly without any fatigue or memory of the test. In addition, the range of certainty around the scores should be small enough to support educational decisions. The reliability/precision of the 2023 NSCAS assessments was examined through analysis of measurement error under simulated and operational conditions, as follows:

- Score precision and reliability of the Cadabra adaptive constraint-based engine (see Score Precision and Reliability)
- Marginal reliability
- Conditional standard error of measurement (CSEM)
- Cronbach’s alpha and standard error of measurement (SEM) for fixed forms
- Classification accuracy

Combined, these data provide several ways of looking at the reliability of the NSCAS assessments. Simulation results and marginal reliability statistics, as well as Cronbach’s alpha and SEM for the science fixed forms, operate at the content level and provide estimates of reliability for student scores on a test. CSEM and classification accuracy provide important information related to the NSCAS achievement level classifications. These are of particular interest in the context of state accountability requirements.

9.1. Marginal Reliability

Marginal reliability is typically used in adaptive assessments to investigate score stability and is estimated as the ratio of the mean true score variance (i.e., observed score variance minus mean error variance) to observed score variance, as explained in Evaluation Criteria. Table 9.1 presents the marginal reliabilities of scale scores by grade and reporting category for ELA, mathematics, and science. Marginal reliability estimates for the total scores are all at or above 0.80 (the ELA and mathematics estimates are all 0.85 and higher), which is typically considered the minimal acceptable level of reliability. Because reliability for reporting categories is based on fewer items, items have lower reliability than total scores. Appendix E: Marginal Reliability by Demographics provides marginal reliability estimates for the total scores by demographic subgroup.

Table 9.1. Marginal Reliability of Scale Scores

Content Area	Grade	N	Total Score	Reporting Category			
				1	2	3	4
ELA	3	23,260	0.91	0.72	0.75	0.53	0.65
	4	22,918	0.91	0.71	0.73	0.43	0.67
	5	22,977	0.90	0.69	0.73	0.49	0.64
	6	22,851	0.90	0.69	0.73	0.45	0.62
	7	23,430	0.89	0.66	0.72	0.49	0.62
	8	23,886	0.89	0.68	0.71	0.41	0.62

Content Area	Grade	N	Total Score	Reporting Category			
				1	2	3	4
Mathematics	3	23,197	0.95	0.86	0.75	0.80	0.77
	4	22,842	0.95	0.85	0.80	0.80	0.76
	5	22,917	0.95	0.85	0.79	0.78	0.72
	6	22,774	0.95	0.81	0.86	0.72	0.76
	7	23,348	0.95	0.76	0.84	0.76	0.79
	8	23,787	0.95	0.79	0.82	0.84	0.73
Science	5	22,888	0.87	–	–	–	–
	8	23,807	0.85	–	–	–	–

Note. ELA: 1 = Reading Prose and Poetry, 2 = Reading Informational Text, 3 = Vocabulary, 4 = Writing and Foundations of Writing; Mathematics: 1 = Number, 2 = Algebra, 3 = Geometry, 4 = Data; Science: No reporting category.

As shown in Table 9.2, reliability varies by score level (i.e., decile). Observed variance is from the total score, and error variance is calculated for each decile. All students take the same number of items, but the information delivered by the items differs. The most information (and, hence, lower error and higher reliability) is found where the pool has the most items. The NSCAS item pools have more items in the middle than at both ends and are easy relative to the population, resulting in lower reliability with higher scores (Deciles 9 and 10).

Table 9.2. Marginal Reliability—Variance

Content Area	Grade	N	Variance	MSE	Overall	Deciles									
						1	2	3	4	5	6	7	8	9	10
ELA	3	23,260	8239.33	758.17	0.91	0.88	0.90	0.91	0.92	0.92	0.92	0.92	0.92	0.91	0.88
	4	22,918	8536.57	786.20	0.91	0.90	0.91	0.92	0.92	0.92	0.92	0.92	0.91	0.90	0.87
	5	22,977	7742.95	748.17	0.90	0.89	0.91	0.91	0.91	0.91	0.91	0.91	0.91	0.90	0.86
	6	22,851	6653.62	692.33	0.90	0.87	0.89	0.90	0.91	0.91	0.91	0.91	0.91	0.90	0.87
	7	23,430	6703.55	707.99	0.89	0.87	0.89	0.90	0.90	0.90	0.91	0.91	0.90	0.90	0.87
	8	23,886	6413.87	682.02	0.89	0.88	0.89	0.90	0.90	0.91	0.90	0.90	0.90	0.89	0.86
Mathematics	3	23,197	7789.48	386.34	0.95	0.95	0.95	0.95	0.96	0.96	0.95	0.95	0.95	0.95	0.94
	4	22,842	7599.86	374.69	0.95	0.94	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.94
	5	22,917	6960.49	366.23	0.95	0.94	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.94
	6	22,774	7306.57	364.92	0.95	0.94	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
	7	23,348	6993.72	373.72	0.95	0.93	0.94	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
	8	23,787	7776.71	389.70	0.95	0.93	0.94	0.95	0.95	0.95	0.95	0.96	0.96	0.96	0.95
Science	5	22,888	778.11	98.10	0.87	0.87	0.90	0.92	0.92	0.92	0.90	0.90	0.88	0.86	0.70
	8	23,807	912.45	138.98	0.85	0.83	0.88	0.89	0.89	0.89	0.87	0.87	0.84	0.83	0.68

9.2. Conditional Standard Error of Measurement (CSEM)

The conditional standard error of measurement (CSEM) represents the degree of measurement error, in scale score units, and is conditioned on the ability of the student, meaning that the test has different levels of error at different points along the ability scale. When applied to an adaptive assessment, the CSEM will vary for the same scale score. It is therefore necessary to report averages.

CSEMs are especially useful for characterizing measurement precision regarding score levels used for decision-making, such as the cut score that determines student proficiency on an assessment. Table 9.3 presents the CSEMs for the achievement level cut scores that demark proficiency on the NSCAS tests, including the number of students ± 10 scale score points from the cut scores, the mean CSEMs of students near the cut, and the standard deviation (SD) of the CSEMs.

Table 9.3. CSEMs at the Proficient Cut Scores

Content Area	Grade	Level 3/Level 2 Cut Score			Level 2/Level 1 Cut Score		
		N	Mean CSEM	SD	N	Mean CSEM	SD
ELA	3	1,989	25.8	1.0	1,838	26.5	1.3
	4	2,111	26.0	1.0	1,818	27.8	1.0
	5	1,987	26.0	1.1	1,643	26.9	1.2
	6	2,153	24.9	0.8	1,760	25.8	1.0
	7	2,319	25.2	0.8	1,798	26.2	1.5
	8	2,424	24.7	1.1	1,764	26.4	1.4
Mathematics	3	2,267	18.5	0.7	912	19.7	0.9
	4	2,055	18.8	0.7	1,000	19.1	0.7
	5	2,280	18.9	0.6	1,416	18.6	0.8
	6	2,537	18.8	0.8	1,402	18.5	0.7
	7	2,365	19.2	0.7	1,495	18.6	0.8
	8	2,326	19.4	0.8	1,648	18.5	0.7
Science	5	6,111	8.1	0.3	4,049	10.7	0.8
	8	6,192	10.0	0.0	2,885	13.3	0.4

Note. Level 3 = *Developing*, Level 2 = *On Track*, and Level 1 = *Advanced*.

Table 9.4 presents the overall and by-decile CSEM. The overall CSEM is slightly higher for ELA (from 26.0 to 27.9) than for mathematics (from 19.1 to 19.7), which is expected due to the different conversion slopes. The low CSEM for science is expected, as its conversion slope is smaller than those of ELA or mathematics. CSEM is also relatively similar between Deciles 2 and 9, while the CSEM tends to be higher at the first and last decile. This suggests that item pools have more items in the middle than at both ends and that more difficult items are needed for both ELA and mathematics, which is consistent with reliability results. Appendix F: Scatterplots for Scale Score CSEM presents scatterplots for scale score CSEMs by reporting category for each content area and grade.

Table 9.4. Mean CSEMs by Decile

Content Area	Grade	Mean CSEM	Mean CSEM by Decile									
			1	2	3	4	5	6	7	8	9	10
ELA	3	27.4	31.3	28.3	27.0	26.0	25.7	25.6	25.8	26.2	27.5	30.8
	4	27.9	29.8	27.8	26.9	26.3	26.0	25.9	26.4	27.4	28.9	33.7
	5	27.2	28.8	26.3	26.1	26.2	26.0	25.9	26.1	26.6	27.7	32.4
	6	26.2	29.9	27.1	25.6	25.1	24.9	24.7	24.7	25.1	26.2	29.0
	7	26.5	29.8	27.1	25.9	25.4	25.3	25.2	25.2	25.5	26.3	29.5
	8	26.0	28.1	26.1	25.4	24.8	24.5	24.7	25.1	25.6	26.5	29.5
Mathematics	3	19.6	20.5	19.7	19.0	18.6	18.6	18.9	19.3	19.4	19.7	22.2
	4	19.3	20.8	19.2	18.6	18.6	18.9	19.1	19.2	19.2	19.1	20.4
	5	19.1	20.1	18.8	18.9	19.0	18.6	18.5	18.6	18.7	18.7	20.9
	6	19.1	20.7	19.7	18.9	18.8	18.8	18.8	18.8	18.6	18.5	19.2
	7	19.3	21.7	19.9	19.5	19.2	18.8	18.7	18.6	18.6	18.7	19.3
	8	19.7	23.0	21.0	20.2	19.6	19.3	18.9	18.7	18.5	18.5	19.1
Science	5	9.6	9.9	8.6	8.0	8.0	8.0	9.0	9.0	9.5	10.5	14.5
	8	11.6	12.4	10.6	10.0	10.0	10.0	11.0	11.0	12.0	12.6	16.7

9.3. Classification Accuracy

Classification accuracy is a measure of how accurately test scores place students into reporting category levels. It refers to the agreement between the actual classifications using observed cut scores and true classifications based on known true cut scores. It is common to estimate classification accuracy by using a psychometric model to find true scores corresponding to observed scores. The likelihood of inaccurate placement depends on the amount of error associated with scores, especially those nearest cut points.

Classification accuracy was calculated as follows (SBAC, 2016):

1. For each student, a normal distribution was constructed, with means equal to the scale score estimate and standard deviation equal to the SEM as a plausible true score distribution.
2. For each student, the proportion of that normal distribution that fell within each achievement level was calculated.
3. Within the groups of students assigned to a particular achievement level (Level 3, 2, or 1 for the overall score), the sums of the proportions over students were computed. This provided estimates of the number of students whose true score falls within a level for each assigned achievement level. These sums were then expressed as a proportion of the total sample (i.e., expected proportion).
4. With the table of expected proportions, correct classification rates were then defined. This is the proportion of students whose true classification agrees with the assigned level among the subset of students with that assigned level.
5. The overall classification rate is the sum of the proportions of students whose true score level agrees with the assigned level divided by the total proportion of students assigned to a level.

Table 9.5 presents the classification accuracy results by content area, grade, and achievement level. Overall, classification accuracy ranges from 0.827 (ELA grade 4) to 0.902 (mathematics grade 4). In general, classification accuracy is moderate to high. Considering that the magnitude of classification accuracy is influenced by key features of test design (including the number of items, number of cut scores, and the reliability and associated SEM), the classification accuracy suggests that accurate level classifications are being made for Nebraska students on the NSCAS assessments.

Table 9.5. Classification Accuracy by Achievement Level

Grade	Achievement Level	N	%	Expected Proportion ^a			Class. Acc.	Overall Class. Acc.
				L3	L2	L1		
ELA								
3	Developing	8,766	0.38	0.34	0.04	0.00	0.902	0.841
	On Track	9,428	0.41	0.05	0.32	0.04	0.788	
	Advanced	5,066	0.22	0.00	0.04	0.18	0.835	
4	Developing	10,306	0.45	0.41	0.04	0.00	0.909	0.827
	On Track	7,573	0.33	0.05	0.24	0.04	0.724	
	Advanced	5,039	0.22	0.00	0.04	0.18	0.814	
5	Developing	9,917	0.43	0.39	0.04	0.00	0.912	0.837
	On Track	8,312	0.36	0.05	0.27	0.04	0.746	
	Advanced	4,748	0.21	0.00	0.03	0.17	0.836	

Grade	Achievement Level	N	%	Expected Proportion ^a			Class. Acc.	Overall Class. Acc.
				L3	L2	L1		
6	Developing	10,204	0.45	0.41	0.04	0.00	0.911	0.833
	On Track	8,622	0.38	0.05	0.28	0.04	0.753	
	Advanced	4,025	0.18	0.00	0.03	0.14	0.807	
7	Developing	10,723	0.46	0.42	0.04	0.00	0.906	0.833
	On Track	9,069	0.39	0.05	0.29	0.04	0.760	
	Advanced	3,638	0.16	0.00	0.03	0.12	0.800	
8	Developing	8,720	0.37	0.32	0.04	0.00	0.888	0.837
	On Track	11,424	0.48	0.05	0.39	0.04	0.810	
	Advanced	3,742	0.16	0.00	0.03	0.13	0.803	
Mathematics								
3	Developing	9,695	0.42	0.39	0.03	0.00	0.921	0.898
	On Track	10,452	0.45	0.04	0.40	0.02	0.880	
	Advanced	3,050	0.13	0.00	0.02	0.12	0.885	
4	Developing	9,585	0.42	0.39	0.03	0.00	0.933	0.902
	On Track	10,500	0.46	0.04	0.41	0.02	0.880	
	Advanced	2,757	0.12	0.00	0.02	0.11	0.868	
5	Developing	7,939	0.35	0.32	0.03	0.00	0.913	0.888
	On Track	10,659	0.47	0.04	0.40	0.02	0.869	
	Advanced	4,319	0.19	0.00	0.02	0.17	0.894	
6	Developing	9,641	0.42	0.39	0.04	0.00	0.915	0.882
	On Track	9,114	0.40	0.04	0.34	0.02	0.845	
	Advanced	4,019	0.18	0.00	0.02	0.16	0.892	
7	Developing	7,932	0.34	0.31	0.03	0.00	0.906	0.881
	On Track	10,736	0.46	0.04	0.39	0.03	0.854	
	Advanced	4,680	0.20	0.00	0.02	0.18	0.900	
8	Developing	9,064	0.38	0.35	0.03	0.00	0.916	0.880
	On Track	9,214	0.39	0.04	0.32	0.03	0.835	
	Advanced	5,509	0.23	0.00	0.02	0.21	0.897	
Science								
5	Developing	5,248	0.23	0.20	0.03	0.00	0.891	0.857
	On Track	13,985	0.61	0.05	0.53	0.03	0.874	
	Advanced	3,655	0.16	0.00	0.04	0.12	0.744	
8	Developing	8,305	0.35	0.31	0.04	0.00	0.888	0.848
	On Track	13,480	0.57	0.05	0.47	0.04	0.832	
	Advanced	2,022	0.09	0.00	0.02	0.07	0.788	

^a L3 = *Developing*, L2 = *On Track*, and L1 = *Advanced*

9.4. Reliability for Fixed Forms (Science)

Cronbach's alpha reliability coefficient is a frequently used measure of internal consistency of the responses to a set of items measuring an underlying, unidimensional trait. Reliability coefficient alpha expresses the consistency of test scores as the ratio of true score variance to total score (observed) variance (true score variance + error variance). A larger index would indicate that test scores were less influenced by random sources of error. The reliability coefficient is a "unitless" index, which can be compared from test to test and ranges from 0.0 to 1.0, where 0.80 is typically considered the minimally acceptable level of reliability for assessments such as NSCAS. While sensitive to random error associated with content sampling variability, the index is not sensitive to other types of errors, such as temporal stability

or variability in performance that might occur across different testing occasions. Cronbach's alpha is computed as follows (Crocker & Algina, 1986):

$$\hat{\alpha} = \frac{k}{k-1} \left(1 - \frac{\sum \sigma_j^2}{\sigma_X^2} \right)$$

where k is the number of items, σ_X^2 is the total score variance, and σ_j^2 is the variance of item j .

The SEM is an index of the random variability in test scores in raw score units and is defined as follows:

$$SEM = SD\sqrt{1 - \hat{\alpha}}$$

where SD represents the standard deviation of the raw score distribution, and $\hat{\alpha}$ represents Cronbach's alpha. The overall SEM is expressed in raw score units and is a test-level statistic. Table 9.6 presents Cronbach's alpha reliability coefficients by demographics for the science fixed forms, along with the SEMs. The alpha reliability coefficients are similar to marginal reliability (reported in Table 9.1 and Table 9.2).

Table 9.6. Cronbach's Alpha (Internal Consistency) by Demographics for Science Fixed Forms

Grade	Demographic Group ^a		#Items	Reliability	SEM
5	Grade 5 Overall		21	0.67	16.02
	Gender	Female	21	0.65	15.65
		Male	21	0.69	16.22
	Ethnicity	AI/AN	21	0.51	16.07
		Asian	21	0.67	17.08
		Black or African American	21	0.61	15.80
		Hispanic	21	0.61	15.12
		NH/PI	21	0.67	16.25
		White	21	0.66	15.71
		Two or More Races	21	0.65	16.15
	FRL	Yes	21	0.63	15.51
		No	21	0.66	15.70
	LEP	Yes	21	0.57	15.28
		No	21	0.67	15.88
	SPED	Yes	21	0.60	16.31
		No	21	0.66	15.50
8	Grade 8 Overall		27	0.75	15.10
	Gender	Female	27	0.73	14.98
		Male	27	0.77	15.09
	Ethnicity	AI/AN	27	0.73	14.05
		Asian	27	0.77	16.03
		Black or African American	27	0.70	14.32
		Hispanic	27	0.72	14.44
		NH/PI	27	0.77	14.98

Grade	Demographic Group ^a		#Items	Reliability	SEM
		White	27	0.72	15.12
		Two or More Races	27	0.75	14.76
	FRL	Yes	27	0.73	14.71
		No	27	0.72	15.19
	LEP	Yes	27	0.65	14.00
		No	27	0.74	15.05
	SPED	Yes	27	0.69	14.56
		No	27	0.73	15.01

^a AI/AN = American Indian or Alaskan Native; NH/PI = Native Hawaiian or Other Pacific Islander; FRL = free and reduced lunch; LEP = limited English proficient; SPED = special education

Section 10: Validity

The *Standards for Educational and Psychological Testing* refers to validity as the “degree to which evidence and theory support the interpretations of test scores for proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing and evaluating tests” (AERA et al., 2014, p. 11). Validating a test score interpretation is not a quantifiable property but an ongoing process, beginning at initial conceptualization of the construct and continuing throughout the entire assessment process. Every aspect of an assessment development and administration process provides evidence in support of (or a challenge to) the validity of the intended inferences about what students know based on their score, including design, content specifications, item development, test constraints, psychometric quality, standard setting, and administration.

This technical report covers the different phases of the testing cycle and provides different pieces of technical quality evidence along the way. It provides relevant evidence and a rationale in support of test-score interpretations and intended uses based on the *Standards*, considered to be “the most authoritative statement of professional consensus regarding the development and evaluation of educational and psychological tests” (Linn, 2006, p. 27). The validity argument begins with a statement of the assessment’s intended purposes, followed by the evidentiary framework, where available validity evidence is provided to support the argument that the test actually measures what it purports to measure (SBAC, 2016).

While NSCAS assessments offer the additional benefit of reporting category scores that indicate directions for gaining further instructional information through the interim system or classroom observation, scores based on NSCAS are as equally reliable and valid as a traditional end-of-year assessment due to the following factors: First, NSCAS assessments go through the same rigorous psychometric analyses (such as test reliability, classification accuracy, CSEMs, test information, DIF, and convergent validity check), and the analysis results so far strongly support the reliability and validity claims of the NSCAS assessments. In addition, the test-development process ensures validity of the intended test-score interpretations provided through the Reporting ALDs and scale scores. Last but not least, NSCAS assessments are aligned to grade-level content, and their test scores are suitable for use in accountability systems as a result of a robust development process of table of specifications (TOS), passage and item specifications, and achievement level descriptors (ALDs).

10.1. Intended Purposes and Uses of Test Scores

Building a validity argument begins with identifying the purposes of the assessment and the intended uses of its test scores. The purposes of the NSCAS Growth assessments are as follows:

1. To measure and report Nebraska students’ depth of achievement regarding the Nebraska College and Career Ready Standards
2. To determine if student achievement demonstrates sufficient academic proficiency to be on track for achieving college readiness
3. To measure students’ annual progress toward college and career readiness
4. To inform teachers how student thinking differs along different areas of the scale, as represented by the ALDs, as information to support instructional planning
5. To assess students’ construct-relevant achievement in ELA, mathematics, and science for all students and subgroups of students

As the *Standards* notes, “validation is the joint responsibility of the test developer and the test user. . . . The test user is ultimately responsible for evaluating the evidence in the particular setting in which the test is to be used” (AERA et al., 2014, p. 13). This report provides information about test content and technical quality but does not interfere in the use of scores. Ultimate use of test scores is determined by Nebraska educators. However, some intended uses of the NSCAS test results include the following:

- To supplement teachers’ observations and classroom assessment data and to improve the decisions teachers make about sequencing instructional goals, designing instructional materials, and selecting instructional approaches for groups and individuals
- To identify individuals for summer school and other remediation programs
- To gauge and improve the quality of education at the class, school, system, and state levels throughout Nebraska
- To assess the performance of a teacher, school, or system in conjunction with other sources of information

Unintended uses of the NSCAS include:

- To place students in special-education classes
- To apply group differences in test scores to admission and class grouping
- To narrow a school’s curriculum to exclude learning of objectives that are not assessed

10.2. Sources of Validity Evidence

The *Standards* describes validation as a process of constructing and evaluating arguments for the intended interpretation and use of test scores:

“A sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses. . . .

Ultimately, the validity of an intended interpretation of test scores relies on all the available evidence relevant to the technical quality of a testing system” (AERA et al., 2014, pp. 21–22).

The *Standards* (AERA et al., 2014, pp. 13–19) outlines the following five main sources of validity evidence:

- Evidence based on test content
- Evidence based on response processes
- Evidence based on internal structure
- Evidence based on relations to other variables
- Evidence based on validity and consequences of testing

Evidence based on test content refers to traditional forms of content validity or content-related evidence. Evidence based on response processes refers to the cognitive process engaged in by students when answering test items, or the “evidence concerning the fit between the construct and the detailed nature of performance or response actually engaged in by examinees” (AERA

et al., 2014, p. 15). Evidence based on internal structure refers to the psychometric analyses of “the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based” (AERA et al., 2014, p. 16). Evidence based on relations to other variables refers to traditional forms of criterion-related validity evidence, such as predictive and concurrent validity. Evidence based on validity and consequences of testing refers to the evaluation of the intended and unintended consequences associated with a testing program.

10.3. Evidentiary Validity Framework

Table 10.1 presents an overview of the validity components covered in this technical report.

Table 10.1. Sources of Validity Evidence for Each NSCAS Test Purpose

Test Purpose	Sources of Validity Evidence			
	Test Content	Response Processes	Internal Structure	Relations to Other Variables
1. Measure and report Nebraska students' depth of achievement regarding the Nebraska College and Career Ready Standards.	✓	✓	✓	✓
2. Determine if student achievement demonstrates sufficient academic proficiency to be on track for achieving college readiness.	✓	✓	✓	
3. Measure students' annual progress toward college and career readiness.	✓	✓	✓	
4. Inform teachers how student thinking differs along different areas of the scale, as represented by the ALDs, as information to support instructional planning.	✓	✓	✓	
5. Assess students' construct-relevant achievement in ELA, mathematics, and science for all students and subgroups of students.	✓	✓	✓	

Table 10.2–Table 10.5 examine the types of evidence available for each intended purpose of the NSCAS assessments.

Table 10.2. Sources of Validity Evidence Based on Test Content

Test Purpose	Summary of Evidence	Tech Report Section(s)
1. Measure and report Nebraska students' depth of achievement regarding the Nebraska College and Career Ready Standards.	<ul style="list-style-type: none"> • Bias is minimized through Universal Design and accessibility resources. • Blueprint, passage specifications, and item specifications are aligned to grade-level content, process skills, and associated cognitive complexity. • The item pool and item-selection procedures adequately support the test design. 	2, 9

Test Purpose	Summary of Evidence	Tech Report Section(s)
2. Determine if student achievement demonstrates sufficient academic proficiency to be on track for achieving college readiness.	<ul style="list-style-type: none"> Nebraska's College and Career Ready Standards are based on skills leading to college and career readiness across grades. Blueprint, passage specifications, and item specifications are aligned to grade-level content, process skills, and associated cognitive complexity. 	2
3. Measure students' annual progress toward college and career readiness.	<ul style="list-style-type: none"> Nebraska's College and Career Ready Standards are based on skills leading to college and career readiness across grades. Blueprint, passage specifications and item specifications are aligned to grade-level content, process skills, and associated cognitive complexity. 	2
4. Inform teachers how student thinking differs along different areas of the scale, as represented by the ALDs, as information to support instructional planning.	<ul style="list-style-type: none"> Blueprint, passage specifications, and item specifications are aligned to grade level content, process skills, and associated cognitive complexity. Blueprint and ALDs were developed in consultation with Nebraska educators. Reporting categories align with the structure of the Nebraska standards to support the interpretation of the test results. 	2, 4, 7
5. Assess students' construct-relevant achievement in ELA, mathematics, and science for all students and subgroups of students.	<ul style="list-style-type: none"> Bias is minimized through Universal Design and accessibility resources. Assessments are administered with appropriate accommodations. 	2, 3, 6, 9

Table 10.3. Sources of Validity Evidence Based on Response Processes

Test Purpose	Summary of Evidence	Tech Report Section(s)
1. Measure and report Nebraska students' depth of achievement regarding the Nebraska College and Career Ready Standards.	<ul style="list-style-type: none"> Bias is minimized through Universal Design and accessibility resources. Blueprint, passage specifications, and item specifications are aligned to grade-level content, process skills, and associated cognitive complexity. Achievement levels were set to be consistent with best practices. 	2
2. Determine if student achievement demonstrates sufficient academic proficiency to be on track for achieving college readiness.	<ul style="list-style-type: none"> Blueprint, passage specifications, and item specifications are aligned to grade-level content, process skills, and associated cognitive complexity. Achievement levels are vertically articulated. 	2
3. Measure students' annual progress toward college and career readiness.	<ul style="list-style-type: none"> Blueprint, passage specifications, and item specifications are aligned to grade-level content, process skills, and associated cognitive complexity. Achievement levels are vertically articulated. 	2

Test Purpose	Summary of Evidence	Tech Report Section(s)
4. Inform teachers how student thinking differs along different areas of the scale, as represented by the ALDs, as information to support instructional planning.	<ul style="list-style-type: none"> • Blueprint, passage specifications, and item specifications are aligned to grade-level content, process skills, and associated cognitive complexity. • Range and Policy ALDs were developed in consultation with committees of Nebraska educators with the goal of providing information to all Nebraska educators. 	2
5. Assess students' construct-relevant achievement in ELA, mathematics, and science for all students and subgroups of students.	<ul style="list-style-type: none"> • Bias is minimized through Universal Design and accessibility resources. • Assessments are administered with appropriate accommodations. 	2, 3, 6, 9

Table 10.4. Sources of Validity Evidence Based on Internal Structure

Test Purpose	Summary of Evidence	Tech Report Section(s)
1. Measure and report Nebraska students' depth of achievement regarding the Nebraska College and Career Ready Standards.	<ul style="list-style-type: none"> • The assessment supports precise measurement and consistent classification. • Achievement levels were set to be consistent with best practices. 	6, 8, 9
2. Determine if student achievement demonstrates sufficient academic proficiency to be on track for achieving college readiness.	<ul style="list-style-type: none"> • Scale is vertically articulated. • Achievement levels were vertically articulated. 	6, 7
3. Measure students' annual progress toward college and career readiness.	<ul style="list-style-type: none"> • The assessment supports precise measurement and consistent classification to support analysis and reporting of longitudinal data. • Scale is vertically articulated. • Achievement levels are vertically articulated. 	6, 7, 9
4. Inform teachers how student thinking differs along different areas of the scale, as represented by the ALDs, as information to support instructional planning.	<ul style="list-style-type: none"> • Range and Policy ALDs were developed in consultation with committees of Nebraska educators with the goal of providing information to all Nebraska educators. • Reporting categories align with the structure of the Nebraska standards to support the interpretation of the test results. • Items are aligned with ALDs in order to support item-writing processes. 	2, 7
5. Assess students' construct-relevant achievement in ELA, mathematics, and science for all students	<ul style="list-style-type: none"> • The assessment supports precise measurement and consistent classification for all students. • DIF analysis was completed for all items across all required subgroups. 	6, 9

Test Purpose	Summary of Evidence	Tech Report Section(s)
and subgroups of students.		

Table 10.5. Sources of Validity Evidence Based on Relations to Other Variables

Test Purpose	Summary of Evidence	Tech Report Section(s)
1. Measure and report Nebraska students' depth of achievement regarding the Nebraska College and Career Ready Standards.	<ul style="list-style-type: none"> • Correlations with MAP Growth are high. 	8
2. Determine if student achievement demonstrates sufficient academic proficiency to be on track for achieving college readiness.	<ul style="list-style-type: none"> • No evidence is provided. 	N/A
3. Measure students' annual progress toward college and career readiness.	<ul style="list-style-type: none"> • No evidence is provided. 	N/A
4. Inform teachers how student thinking differs along different areas of the scale, as represented by the ALDs, as information to support instructional planning.	<ul style="list-style-type: none"> • No evidence is provided. 	N/A
5. Assess students' construct-relevant achievement in ELA, mathematics, and science for all students and subgroups of students.	<ul style="list-style-type: none"> • No evidence is provided. 	N/A

10.4. Interpretive Argument Claims

The test scores for the 2023 NSCAS assessments support their intended purpose, and the interpretation of the test scores, after the careful development of the Reporting ALDs, support that the test scores describe where the students are in their learning at the end of the year based on the Nebraska College and Career Ready Standards. The claims to support this are documented in this technical report, as shown in Table 10.6.

Table 10.6. Interpretive Argument Claims—Evidence to Support Essential Validity Elements

Argument	Tech Report Section(s)	Evidence
Careful test and item development occurred to ensure that the test measured the College and Career Ready Standards.	2. Test Design and Development	Description of the development and review processes for items, passages, and tests
Test score interpretations are comparable across students.	6. Psychometric Analyses 9. Reliability	Simulations, analyses of test information, conditional standard errors of measurement, classification accuracy, and reliability estimates; blueprint comparability across students; item analyses, calibration and linking procedures
Test administrations were secure and standardized.	3. Test Administration and Security	Test-administration procedures, including administration training, test accommodations, test security, and availability of help desk during testing window
Scoring was standardized and accurate.	4. Scoring and Reporting	Scoring rules and procedures; quality control of operational scoring
Achievement standards were rigorous and technically sound.	7. Standard Setting	Documentation of the mathematics standard setting procedures, ELA cut score review process, and the science standard setting procedures, including the methodology, identification of workshop participants, and implementation process, as well as ALD development and validation
Assessments were accessible to all students and fair across student subgroups.	3. Test Administration and Security 6. Psychometric Analyses	Accommodation policy and implementation, sensitivity review, availability of translations, and DIF analyses

10.5. NSCAS Validity Argument

The test development and technical quality of the 2022–2023 NSCAS Growth assessments support the intended test-score interpretations that are provided through the Reporting ALDs and scale scores. The table of specifications (TOS), passage specifications, item specifications, and ALD development process show that the NSCAS assessments are aligned to grade-level content. For ELA and mathematics, there is evidence that the student response processes associated with cognitive complexity specified in the standards and TOS is behaving as intended. As an added dimension for adaptive testing, the NSCAS ELA and mathematics assessments demonstrated that the tests administered to students conform to the blueprints during the adaptive constraint-based engine simulation studies.

The item pool and item-selection procedures used for the adaptive administration adequately support the test designs and blueprints. Content experts developed expanded item types that allow response processes to reveal skills and knowledge. All items were carefully reviewed through multiple cycles of the item-development process for ambiguity, bias, sensitivity,

irrelevant clues, and inaccuracy to ensure the fit between the construct and the nature of performance.

Studies for evidence based on consequences of testing have not been included within the scope of work undertaken to date by NWEA. This evidence may be added in future studies, such as evaluation of the effects of testing on instruction, evaluation of the effects of testing on issues such as high school dropout rates, analyses of students' opportunity to learn, and analyses of changes in textbooks and instructional approaches (SBAC, 2016). The evaluation of unintended consequences may include changes in instruction, diminished morale among teachers and students, increased pressure on students, leading to increased dropout rates, or the pursuit of college majors and careers that are less challenging (SBAC, 2016).

Teacher surveys or focus groups can be used to collect information regarding the use of the tests and how the tests impacted the curriculum and instruction. A better understanding of the extent to which performance gains on assessments reflect improved instruction and student learning (rather than more superficial interventions such as narrow test-preparation activities) would also provide evidence based on consequences of test use. Longitudinal test data, along with additional information collected from Nebraska educators (e.g., information on understanding of learning standards, motivation and effort to adapt the curriculum and instruction to content standards, instructional practices, classroom assessment format and content, use and nature of test assessment preparation activities, professional development), would allow for meaningful analyses and interpretations of the score gain and uniformity of standards, learning expectations, and consequences for all students.

References


- Achieve, Inc. (2018). *Criteria for procuring and evaluating high-quality and aligned summative science assessments*. Retrieved from <https://www.nextgenscience.org/sites/default/files/Criteria03202018.pdf>
- American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. AERA.
- Crocker, L. M., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart, and Winston.
- Dorans, N. J., & Schmitt, A. P. (1991). *Constructed response and differential item functioning: A pragmatic approach* (Research Report No. ETS-RR-91-47). Educational Testing Service (ETS). <https://doi.org/10.1002/j.2333-8504.1991.tb01414.x>
- EdMetric. (2019). *Alignment study for Nebraska Student-Centered Assessment System, mathematics grades 3–8*. Report provided to NDE.
- Egan, K. L., Schneider, M. C., & Ferrara, S. (2012). Performance level descriptors: History, practice and a proposed framework. In G. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 79–106). Routledge.
- Elementary and Secondary Education Act (ESEA) as amended by the ESSA. 20 U.S.C. § 6301 (1965). <https://www2.ed.gov/documents/essa-act-of-1965.pdf>
- French, A. W., & Miller, T. R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement*, 33(3), 315–332. <https://www.jstor.org/stable/1435375>
- Fu, J., & Monfils, L. (2016). *LDIF_ES: A SAS macro for logistic regression tests for differential item functioning of dichotomous and polytomous items*. (Research Memorandum ETS RM-16-17). Princeton, NJ: Educational Testing Service (ETS). <https://www.ets.org/Media/Research/pdf/RM-16-17.pdf>
- Gómez-Benito, J., Hidalgo, M. D., & Padilla, J.-L. (2009). Efficacy of effect size measures in logistic regression: An application for detecting DIF. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 5(1), 18–25. <https://doi.org/10.1027/1614-2241.5.1.18>
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Springer. <https://doi.org/10.1007/978-94-017-1988-9>
- Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer, & H. I. Brown (Eds.), *Test Validity* (pp. 129–145). Lawrence Erlbaum Associates, Inc.
- Huff, K., Warner, Z., & Schweid, J. (2016). Large-scale standards-based assessments of educational achievement. In A. A. Rupp & J. P. Leighton (Eds.), *The Wiley handbook of*

- cognition assessment: Frameworks, methodologies, and applications*, pp. 397–426.
<https://doi.org/10.1002/9781118956588.ch17>
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73. <https://doi.org/10.1111/jedm.12000>
- Lewis, D. M., Mitzel, H. C., Mercado, R. L., & Schulz, E. M. (2012). The Bookmark standard setting procedure. In G. Cizek (Ed.) *Setting performance standards: Foundations, methods, and innovations* (2nd Ed.). Routledge.
- Linacre, J. M. (2015). *Winsteps®* (Version 3.91.0.0) [Computer software]. Portland, Oregon: Winsteps.com. Available from <https://www.winsteps.com/>
- Linn, R. L. (2006). The standards for educational and psychological testing: Guidance in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of Test Development*. Routledge.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149–174.
<https://doi.org/10.1007/BF02296272>
- Messick, S. (1994). *Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning* (Research Report No. RR-94-45). Educational Testing Service (ETS).
<https://files.eric.ed.gov/fulltext/ED380496.pdf>
- National Center for Research on Evaluation, Standards, & Student Testing (CRESST). (2015). *Simulation-based evaluation of the smarter balanced summative assessments*. [Tech. Rep.]. Retrieved from <https://portal.smarterbalanced.org/library/en/simulation-based-evaluation-of-the-smarter-balanced-summative-assessments.pdf>
- National Research Council. (2012). *A framework for K–12 science education: Practices, crosscutting concepts, and core ideas*. Committee on a Conceptual Framework for New K–12 Science Education Standards. Board on Science Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.
- Nebraska Department of Education (NDE). (2023, July). *Nebraska Student-Centered Assessment System (NSCAS) accessibility manual*. <https://www.education.ne.gov/wp-content/uploads/2023/07/NSCAS-Accessibility-Manual-7-12-2023.pdf>
- No Child Left Behind Act of 2002 (NCLB), Pub. L. No. 107–110, § 115, Stat. 1425 (2002).
<https://www.govinfo.gov/content/pkg/BILLS-107hr1enr/pdf/BILLS-107hr1enr.pdf>
- NWEA. (2020). *Constraint-based engine scientific approach and methodology* [Confidential Tech Rep.].
- NWEA. (2021a, October). *Constraint-based engine simulation report for the winter 2021–2022 NSCAS ELA and mathematics assessments* [Confidential Tech Rep.].
- NWEA. (2022a, February). *Constraint-based engine simulation report for the spring 2021–2022 NSCAS ELA and mathematics assessments* Confidential Tech Rep.].

- NWEA. (2022b, May). *Constraint-based engine evaluation report for the winter 2021–2022 NSCAS ELA and mathematics assessments* Confidential Tech Rep.].
- NWEA. (2022c, June). *Constraint-based engine evaluation report for the spring 2021–2022 NSCAS ELA and Mathematics assessments* Confidential Tech Rep.].
- Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist* 51(1), 59–81. <https://doi.org/10.1080/00461520.2016.1145550>
- Phillips, S., & Camara, W. J. (2006). Educational measurement, 733–755.
- Plake, B. S., Huff, K., & Reshetar, R. (2010). Evidence-centered assessment design as a foundation for achievement-level descriptor development and for standard setting. *Applied Measurement in Education*, 23(4), 342–357. <https://doi.org/10.1080/08957347.2010.510964>
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. MESA Press.
- Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (Expanded ed.). University of Chicago Press.
- Samajima, F. (1994). Estimation of reliability coefficients using the test information function and its modifications. *Applied Psychological Measurement*, 18(3), 229–244. <https://doi.org/10.1177/014662169401800304>
- Schneider, M. C., & Johnson, R. L. (2019). *Creating and implementing student learning objectives to support student learning and teacher evaluation*. Taylor and Francis.
- Schneider, M. C., Huff, K. L., Egan, K. L., Gaines, M. L., & Ferrara, S. (2013). Relationships among item cognitive complexity, contextual demands, and item difficulty: Implications for achievement-level descriptors. *Educational Assessment*, 18(2), 99–121. <https://doi.org/10.1080/10627197.2013.789296>
- Smarter Balanced Assessment Consortium (SBAC). (2016). *Smarter Balanced Assessment Consortium: 2014–15 technical report*. Retrieved from <https://portal.smarterbalanced.org/library/en/2014-15-technical-report.pdf>.
- Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27(4), 361–370. <https://www.jstor.org/stable/1434855>
- U.S. Department of Education (USDE). (2018). *A state's guide to the U.S. Department of Education's assessment peer review process*. U.S. Department of Education, Office of Elementary and Secondary Education. <https://www2.ed.gov/admins/lead/account/saa/assessmentpeerreview.pdf>
- Van der Linden, W. J., & Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement*, 22(3), 259–270. <https://doi.org/10.1177/01466216980223006>

- Webb, N. L. (1997). Criteria for alignment of expectations and assessments on mathematics and science education. *Research Monograph No. 6*. Council of Chief State School Officers and National Institute for Science Education (NISE). University of Wisconsin-Madison.
- Webb, N. L. (1999). Alignment of science and mathematics standards and assessments in four states. *Research Monograph No. 18*. Council of Chief State School Officers and National Institute for Science Education (NISE). University of Wisconsin-Madison.
http://archive.wceruw.org/nise/Publications/Research_Monographs/vol18.pdf
- Webb, N. L. (2002). *Alignment study in language arts, mathematics, science, and social studies of state standards and assessments for four states: A study of the state collaborative on assessment & Student standards (SCASS), technical issues in large-scale assessment (TILSA)*. Council of Chief State School Officers.
- Webb, N. L. (2005). *Webb alignment tool: Training manual*. Wisconsin Center of Education Research. University of Wisconsin-Madison. <https://watv2.wceruw.org/>
- Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education*, 20(1), 7–25.
<https://doi.org/10.1080/08957340709336728>
- Wise, S. L., Kingsbury G. G., & Webb, N. L. (2015). Evaluating content alignment in computerized adaptive testing. *Educational Measurement: Issues and Practice*, 34(4), 41–48. <https://doi.org/10.1111/emip.12094>
- Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 14(2), 97–116. <https://www.jstor.org/stable/1434010>
- Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and likert-type (ordinal) item scores*. Directorate of Human Resources Research and Evaluation, Department of National Defense.
- Zwick, R., Donoghue, J. R., & Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement*, 30(3), 233–251.
<https://www.jstor.org/stable/1435045>
- Zwick, R., Thayer, D. T., & Wingersky, M. (1994). DIF analysis for pretest items in computer-adaptive testing. (Research Report No. RR-94-33). Educational Testing Service (ETS).
<https://doi.org/10.1002/j.2333-8504.1994.tb01606.x>

Appendix A: Data Review Cheat Sheet



Data Review Cheat Sheet

Use this document as a guide when reviewing the NSCAS field test items. It includes flagging criteria for four different scenarios:

- General (both multiple-choice and non-multiple-choice items)
- Multiple-choice items
- Non-multiple-choice items (both 1- and 2-point items)
- Non-multiple-choice items (2-point items only)

References starting with “cia,” “fit,” or “dif” are how the statistics are identified in the data review file. The data review file also contains definitions above the statistics to clarify their meaning. A one-page summary of the statistical flags is located at the end of the document.


DIF			
Statistic	Flag	Meaning	Implication for Data Review
DIF of gender or ethnicity	C+ or C-	Item is flagged for potential bias toward a certain group of students.	Is there anything that could trigger the bias toward certain groups of students?

Page 1 of 5

Multiple-Choice Items			
Statistic	Flag	Meaning	Implication for Data Review
P-value Percent of students who got the item correct. (cia_Pval)	< 0.2 or > 0.9	Less than 20% of students got the item correct, or more than 90% of students got it correct.	Does it make sense that an item seems very difficult or very easy?
Option percentages (cia_Pct_Opt1-4)	Distractor % > P-value	More students chose a distractor than the key	Is the answer key accurate? Is the distractor appropriate (common error, etc.)?
Omit (cia_Pct_Omit)	> 5%	More than 5% of students are omitting this item.	Is there anything that could make this item confusing to students?
Item-total correlation aka Point Biserial (cia_ItemTotalCorr)	< 0.2	The item is not differentiating between high- and low-performing students.	Is the answer key accurate?
Item-total correlation for options (cia_ItemTotalCorr_Opt1-4)	> 0.05	An incorrect answer is pulling higher scoring students.	Is there anything that a distractor is doing for high-performing students to select it as an answer? Or is there a possibility for two correct answers? Is the distractor appropriate (common error, etc.)?
IRT Difficulty or Step parameters are extremely High	>=4.25	Probability of getting an item correct may require extremely high ability	Is the item too difficult for even high performing students to get it correct?
Do not use items if items have:			
<ul style="list-style-type: none"> Negative item-total correlation 			

Non-MC Items (Both 1-and 2-point items)			
Statistic	Flag	Meaning	Implication for Data Review
Low student count for each score (cia_Pct_Opt1-3)	= 0	No one got a certain score (e.g., no student got a score of 1).	Is there anything in the item that could cause students to not earn certain scores? Is the key correct?
Item-total correlation (cia_ItemTotalCorr)	< 0.2	The item is not differentiating between high- and low-performing students.	Are the keys accurate? If step parameters are flagged and item total correlation is flagged, the item may not be showing more sophisticated thinking in the content across score points. Is the item asking for the same skill more times?
Item-total correlation for score of 0 (cia_ItemTotalCorr_Opt1)	> 0.0	A score of 0 on the item is not differentiating achievement levels as expected.	Is there a reason earning 0 points is happening more often for high-performing students than low-performing?
Item-total correlation for score of 0 > Item-total correlation for score of 1	cia_ItemTotalCorr_Opt1 > cia_ItemTotalCorr_Opt2	A score of 0 on the item is better differentiating achievement levels than a score of 1.	Is there anything that could make the item perform the opposite of what is expected for high- vs. low-performing students who got a score of 0 vs. 1?
IRT Difficulty or Step parameters are extremely High	>=4.25	Probability of getting an item correct may require extremely high ability	Is the item too difficult for even high performing students to get it correct?
Step parameters [Step 1, Step2]	Step 1 > Step 2	Step parameters are not ordered in value (e.g., the difficulty of score 1 > the difficulty of score 2). There is not a good separation of students into different stages of learning.	Do students have to show more substantive knowledge to earn the second point? Is the same skill being repeated causing the difficulty to stay the same across steps 1 and 2? Is there another reason the difficulty is not increasing across points?
Do not use items if items have:			
<ul style="list-style-type: none"> Negative item-total correlation 			

Non-MC Items (2-point items only)			
Statistic	Flag	Meaning	Implication for Data Review
Item-total correlation for score of 1 > Item-total correlation for score of 2	$\text{cia_ItemTotalCorr_Opt2} > \text{cia_ItemTotalCorr_Opt3}$	A score of 1 on the item is better at differentiating achievement levels than a score of 2.	Is there anything that could make the item perform the opposite of what is expected for high- vs. low-performing students who got a score of 1 vs. 2?
Item-total correlation for score of 2 (cia_ItemTotalCorr_Opt3)	< 0.2	A score of 2 on the item is not differentiating achievement levels as expected.	Is there a reason earning 2 points is happening more often for low-performing students than high-performing?
IRT Difficulty or Step parameters are extremely High	≥ 4.25	Probability of getting an item correct may require extremely high ability	Is the item too difficult for even high performing students to get it correct?
Step parameters [Step 1, Step2]	Step 1 > Step 2	Step parameters are not ordered in value (e.g., the difficulty of score 1 > the difficulty of score 2). There is not a good separation of students into different stages of learning.	Do students have to show more substantive knowledge to earn the second point? Is the same skill being repeated causing the difficulty to stay the same across steps 1 and 2? Is there another reason the difficulty is not increasing across points?
Do not use 2-point items if items have: <ul style="list-style-type: none"> Negative item-total correlation No second-step parameters. 			

		Data Review Cheat Sheet	
	Label	Statistics	Flags
MC items	Pvalue_LOW/ Pvalue_HIGH	P-value	< 0.2 or > 0.9
	Pvalue_Dis	Option percentages	Distractor % > P-value
	Pbis_LOW	Item-total correlation	< 0.20
	Pbis_Dis	Item-total correlation for distractors	> 0.05
Non-MC items (Both 1- and 2-point items)	Pvalue_LOW/ Pvalue_HIGH	P-value	< 0.2 or > 0.9
	N_012	Low student count for each score	= 0
	Pbis_LOW	Item-total correlation	< 0.2
	Score_0_Pbis	Item-total correlation for score of 0	> 0.0
	Score_0Vs1_Pbis	Item-total correlation for score of 0 > item-total correlation for score of 1	
Non-MC items (2-point items only)	Score_1Vs2_Pbis	Item-total correlation for score of 1 > item-total correlation for score of 2	
	Score_2_Pbis	Item-total correlation for score of 2	< 0.2
Item Parameters	itemFlag_IRT_Parameter	IRT Difficulty or Step parameters are extreme	>=4.25
	itemFlag_IRT_ReversedStep	Reversed Step parameters	Step 1 > Step 2
DIF	itemFlag_Gender_DIF/ itemFlag_Black_DIF/ itemFlag_Hispanic_DIF	DIF of gender or ethnicity	C+ or C-
Do not use items if items have: <ul style="list-style-type: none"> Negative item-total correlation No second step parameters 			

Appendix B: Summary of *P* Values by Item TypeTable B.1. Summary of *P* Values by Item Type—Operational Items

Grade	Item Type	#Items	Mean	SD	Min.	Max.	#Items by P Value Range									
							≤ 0.1	≤ 0.2	≤ 0.3	≤ 0.4	≤ 0.5	≤ 0.6	≤ 0.7	≤ 0.8	≤ 0.9	> 0.9
ELA																
3	Choice Multiple	44	0.52	0.12	0.20	0.72	0	0	2	5	11	16	9	1	0	0
	Choice Single	519	0.54	0.14	0.19	0.92	0	1	17	65	127	150	90	52	16	1
	Composite	42	0.42	0.10	0.12	0.67	0	1	3	10	22	4	2	0	0	0
	Gap Match Multiple	31	0.47	0.13	0.20	0.69	0	0	5	3	8	11	4	0	0	0
	Gap Match Single	1	0.50	--	0.50	0.50	0	0	0	0	1	0	0	0	0	0
	Hot Text	2	0.33	0.31	0.11	0.56	0	1	0	0	0	1	0	0	0	0
4	Choice Multiple	63	0.57	0.10	0.35	0.81	0	0	0	3	14	19	22	3	2	0
	Choice Single	419	0.59	0.15	0.20	0.98	0	0	3	38	97	99	93	58	21	10
	Composite	34	0.53	0.09	0.31	0.71	0	0	0	2	14	11	6	1	0	0
	Gap Match Multiple	24	0.52	0.11	0.31	0.74	0	0	0	3	7	8	5	1	0	0
	Gap Match Single	1	0.54	--	0.54	0.54	0	0	0	0	0	1	0	0	0	0
	Hot Text	2	0.49	0.04	0.46	0.51	0	0	0	0	1	1	0	0	0	0
5	Choice Multiple	60	0.58	0.09	0.37	0.78	0	0	0	1	12	24	16	7	0	0
	Choice Single	429	0.57	0.14	0.25	0.95	0	0	8	35	97	121	101	42	20	5
	Composite	23	0.48	0.11	0.28	0.66	0	0	1	4	7	6	5	0	0	0
	Gap Match Multiple	24	0.53	0.18	0.05	0.78	1	0	1	3	5	4	6	4	0	0
	Gap Match Single	3	0.64	0.13	0.49	0.74	0	0	0	0	1	0	1	1	0	0
	Hot Text	7	0.49	0.09	0.39	0.62	0	0	0	1	4	1	1	0	0	0
6	Choice Multiple	51	0.49	0.10	0.22	0.69	0	0	2	8	14	19	8	0	0	0
	Choice Single	461	0.56	0.15	0.13	0.96	0	2	13	43	112	115	92	57	25	2
	Composite	43	0.48	0.12	0.14	0.74	0	2	2	3	16	15	3	2	0	0
	Gap Match Multiple	26	0.51	0.12	0.26	0.72	0	0	1	5	4	10	4	2	0	0
	Gap Match Single	1	0.54	--	0.54	0.54	0	0	0	0	0	1	0	0	0	0
	Hot Text	4	0.45	0.22	0.18	0.67	0	1	0	1	0	0	2	0	0	0
7	Choice Multiple	47	0.50	0.09	0.26	0.70	0	0	1	3	19	19	5	0	0	0
	Choice Single	392	0.56	0.13	0.20	0.95	0	1	9	33	84	123	88	37	14	3
	Composite	37	0.51	0.07	0.39	0.65	0	0	0	3	10	21	3	0	0	0
	Gap Match Multiple	17	0.53	0.13	0.29	0.73	0	0	1	1	5	4	4	2	0	0
	Gap Match Single	4	0.31	0.18	0.10	0.55	0	1	1	1	0	1	0	0	0	0
	Hot Text	9	0.54	0.05	0.49	0.61	0	0	0	0	3	5	1	0	0	0
8	Choice Multiple	52	0.46	0.12	0.08	0.82	1	1	1	9	26	9	4	0	1	0
	Choice Single	440	0.59	0.14	0.29	0.99	0	0	3	28	82	130	109	54	20	14

Appendix B: Summary *P* Values by Item Type

Grade	Item Type	#Items	Mean	SD	Min.	Max.	#Items by <i>P</i> Value Range									
							≤ 0.1	≤ 0.2	≤ 0.3	≤ 0.4	≤ 0.5	≤ 0.6	≤ 0.7	≤ 0.8	≤ 0.9	> 0.9
	Composite	47	0.49	0.11	0.13	0.77	0	1	0	7	19	13	5	2	0	0
	Gap Match Multiple	28	0.51	0.12	0.33	0.79	0	0	0	6	9	7	3	3	0	0
	Gap Match Single	3	0.40	0.14	0.27	0.54	0	0	1	1	0	1	0	0	0	0
	Hot Text	12	0.55	0.11	0.43	0.77	0	0	0	0	6	3	2	1	0	0
Mathematics																
3	Choice Multiple	46	0.53	0.12	0.22	0.81	0	0	1	4	18	11	8	3	1	0
	Choice Single	534	0.50	0.10	0.00	0.84	1	1	1	52	263	148	43	20	5	0
	Composite	56	0.55	0.13	0.18	0.88	0	1	2	3	9	24	12	3	2	0
	Gap Match Multiple	47	0.52	0.10	0.20	0.73	0	1	0	1	19	17	7	2	0	0
	Gap Match Single	6	0.51	0.06	0.42	0.58	0	0	0	0	2	4	0	0	0	0
	Graphic Gap Match	51	0.51	0.12	0.14	0.75	0	2	0	1	23	16	6	3	0	0
	Hot Text	9	0.48	0.08	0.36	0.63	0	0	0	1	5	2	1	0	0	0
	Text Entry	46	0.50	0.11	0.00	0.74	1	0	0	2	20	16	5	2	0	0
4	Choice Multiple	46	0.49	0.08	0.31	0.67	0	0	0	8	17	19	2	0	0	0
	Choice Single	307	0.49	0.07	0.28	0.71	0	0	1	16	164	106	17	3	0	0
	Composite	53	0.50	0.14	0.00	0.75	1	1	3	4	15	18	9	2	0	0
	Gap Match Multiple	31	0.50	0.10	0.36	0.73	0	0	0	4	16	5	4	2	0	0
	Gap Match Single	4	0.59	0.11	0.49	0.70	0	0	0	0	1	1	1	1	0	0
	Graphic Gap Match	43	0.53	0.07	0.39	0.73	0	0	0	1	12	25	3	2	0	0
	Hot Text	17	0.46	0.09	0.30	0.73	0	0	1	4	8	3	0	1	0	0
	Text Entry	57	0.52	0.07	0.34	0.74	0	0	0	2	20	31	3	1	0	0
5	Choice Multiple	48	0.47	0.08	0.34	0.72	0	0	0	10	19	18	0	1	0	0
	Choice Single	366	0.50	0.09	0.00	1.00	1	0	1	27	165	130	37	3	1	1
	Composite	67	0.52	0.12	0.30	0.79	0	0	0	8	25	21	6	7	0	0
	Gap Match Multiple	35	0.49	0.08	0.31	0.69	0	0	0	2	19	11	3	0	0	0
	Gap Match Single	4	0.51	0.18	0.37	0.77	0	0	0	1	2	0	0	1	0	0
	Graphic Gap Match	28	0.50	0.10	0.28	0.69	0	0	1	3	11	9	4	0	0	0
	Hot Text	11	0.49	0.05	0.41	0.60	0	0	0	0	7	3	1	0	0	0
	Text Entry	52	0.52	0.08	0.38	0.78	0	0	0	4	19	22	5	2	0	0
6	Choice Multiple	66	0.44	0.09	0.13	0.72	0	1	3	19	29	11	2	1	0	0
	Choice Single	605	0.50	0.09	0.00	1.00	1	0	6	76	248	210	50	11	2	1
	Composite	62	0.45	0.15	0.00	1.00	1	0	7	12	21	15	5	0	0	1
	Gap Match Multiple	41	0.45	0.10	0.20	0.74	0	0	3	8	21	7	1	1	0	0
	Gap Match Single	1	0.55	--	0.55	0.55	0	0	0	0	0	1	0	0	0	0
	Graphic Gap Match	29	0.48	0.08	0.26	0.63	0	0	1	3	14	9	2	0	0	0
	Hot Text	23	0.44	0.06	0.31	0.60	0	0	0	4	17	2	0	0	0	0
	Text Entry	69	0.51	0.10	0.25	0.85	0	0	1	7	27	26	4	2	2	0
7	Choice Multiple	40	0.40	0.09	0.18	0.61	0	1	3	17	14	3	2	0	0	0

Appendix B: Summary *P* Values by Item Type

Grade	Item Type	#Items	Mean	SD	Min.	Max.	#Items by <i>P</i> Value Range									
							≤ 0.1	≤ 0.2	≤ 0.3	≤ 0.4	≤ 0.5	≤ 0.6	≤ 0.7	≤ 0.8	≤ 0.9	> 0.9
	Choice Single	448	0.48	0.09	0.00	1.00	1	0	5	55	232	133	16	3	0	3
	Composite	48	0.42	0.13	0.14	0.82	0	3	6	11	16	9	2	0	1	0
	Gap Match Multiple	36	0.42	0.08	0.25	0.64	0	0	2	11	18	4	1	0	0	0
	Graphic Gap Match	9	0.43	0.08	0.30	0.54	0	0	0	3	5	1	0	0	0	0
	Hot Text	22	0.41	0.10	0.24	0.54	0	0	3	7	7	5	0	0	0	0
	Text Entry	74	0.47	0.09	0.27	0.75	0	0	4	11	32	24	2	1	0	0
8	Choice Multiple	39	0.40	0.10	0.00	0.52	1	1	5	8	19	5	0	0	0	0
	Choice Single	332	0.47	0.07	0.23	0.69	0	0	6	33	189	97	7	0	0	0
	Composite	55	0.38	0.16	0.00	0.71	1	9	8	9	18	5	4	1	0	0
	Gap Match Multiple	42	0.44	0.08	0.26	0.60	0	0	4	9	21	8	0	0	0	0
	Gap Match Single	3	0.40	0.07	0.32	0.46	0	0	0	1	2	0	0	0	0	0
	Graphic Gap Match	12	0.44	0.08	0.32	0.58	0	0	0	3	7	2	0	0	0	0
	Hot Text	36	0.44	0.11	0.06	0.75	1	0	2	11	11	10	0	1	0	0
	Text Entry	57	0.45	0.09	0.17	0.58	0	2	2	11	23	19	0	0	0	0
Science																
5	Choice Multiple	8	0.55	0.19	0.19	0.83	0	1	0	0	1	3	2	0	1	0
	Choice Single	12	0.62	0.11	0.43	0.79	0	0	0	0	3	1	5	3	0	0
	Composite	4	0.40	0.27	0.13	0.72	0	1	1	0	0	1	0	1	0	0
	Gap Match Multiple	3	0.68	0.30	0.33	0.88	0	0	0	1	0	0	0	0	2	0
	Graphic Gap Match	3	0.70	0.15	0.53	0.81	0	0	0	0	0	1	0	1	1	0
	Hot Text	1	0.46	--	0.46	0.46	0	0	0	0	1	0	0	0	0	0
8	Choice Multiple	3	0.38	0.06	0.32	0.45	0	0	0	2	1	0	0	0	0	0
	Choice Single	15	0.61	0.14	0.29	0.83	0	0	1	0	2	5	3	2	2	0
	Composite	6	0.45	0.23	0.16	0.77	0	1	1	0	1	2	0	1	0	0
	Gap Match Multiple	3	0.59	0.24	0.33	0.79	0	0	0	1	0	0	1	1	0	0
	Graphic Gap Match	3	0.75	0.06	0.68	0.79	0	0	0	0	0	0	1	2	0	0

Table B.2. Summary of *P* Values by Item Type—Field Test Items

Grade	Item Type	#Items	Mean	SD	Min.	Max.	#Items by P Value Range									
							≤ 0.1	≤ 0.2	≤ 0.3	≤ 0.4	≤ 0.5	≤ 0.6	≤ 0.7	≤ 0.8	≤ 0.9	> 0.9
ELA																
3	Choice Multiple	9	0.56	0.12	0.3	0.7	0	0	0	1	2	2	3	1	0	0
	Choice Single	132	0.62	0.16	0.22	0.97	0	0	4	10	20	31	27	20	12	8
	Composite	11	0.36	0.14	0.15	0.56	0	1	4	2	1	3	0	0	0	0
	Gap Match Multiple	5	0.44	0.15	0.27	0.61	0	0	2	0	1	1	1	0	0	0
	Hot Text	4	0.57	0.06	0.51	0.66	0	0	0	0	0	3	1	0	0	0
4	Choice Multiple	8	0.47	0.1	0.36	0.6	0	0	0	2	3	3	0	0	0	0
	Choice Single	97	0.58	0.18	0.03	0.94	1	1	4	9	13	25	18	16	9	1
	Composite	11	0.63	0.12	0.46	0.82	0	0	0	0	2	2	4	2	1	0
	Gap Match Multiple	9	0.52	0.16	0.27	0.81	0	0	1	1	2	3	1	0	1	0
	Hot Text	6	0.47	0.29	0.02	0.82	1	0	0	1	2	0	0	1	1	0
5	Choice Multiple	11	0.62	0.11	0.48	0.78	0	0	0	0	2	4	2	3	0	0
	Choice Single	140	0.6	0.21	0.14	0.99	0	4	8	19	18	19	20	31	13	8
	Composite	11	0.42	0.11	0.26	0.6	0	0	2	3	3	2	1	0	0	0
	Gap Match Multiple	4	0.52	0.08	0.45	0.62	0	0	0	0	2	1	1	0	0	0
	Gap Match Single	1	0.59	--	0.59	0.59	0	0	0	0	0	1	0	0	0	0
	Hot Text	3	0.48	0.24	0.21	0.65	0	0	1	0	0	1	1	0	0	0
6	Choice Multiple	17	0.48	0.15	0.12	0.75	0	1	0	2	8	3	2	1	0	0
	Choice Single	105	0.54	0.18	0.08	0.92	1	1	10	10	18	21	24	11	8	1
	Composite	13	0.41	0.18	0.15	0.75	0	2	1	4	3	1	1	1	0	0
	Gap Match Multiple	3	0.51	0.1	0.41	0.59	0	0	0	0	1	2	0	0	0	0
	Gap Match Single	1	0.22	--	0.22	0.22	0	0	1	0	0	0	0	0	0	0
	Hot Text	2	0.23	0.25	0.05	0.41	1	0	0	0	1	0	0	0	0	0
7	Choice Multiple	21	0.53	0.17	0.06	0.78	1	0	0	3	4	4	5	4	0	0
	Choice Single	98	0.6	0.18	0.13	0.93	0	2	4	8	17	17	19	19	9	3
	Composite	10	0.41	0.12	0.23	0.59	0	0	2	3	3	2	0	0	0	0
	Gap Match Multiple	10	0.26	0.13	0.03	0.53	1	2	4	2	0	1	0	0	0	0
	Gap Match Single	2	0.6	0.37	0.34	0.86	0	0	0	1	0	0	0	0	1	0
	Hot Text	9	0.5	0.17	0.23	0.77	0	0	1	3	0	3	1	1	0	0
8	Choice Multiple	27	0.48	0.22	0	0.78	3	2	1	1	3	8	6	3	0	0
	Choice Single	141	0.52	0.19	0	0.92	2	3	14	23	29	21	21	18	8	2
	Composite	16	0.49	0.16	0.23	0.73	0	0	3	1	4	4	1	3	0	0
	Gap Match Multiple	3	0.43	0.18	0.31	0.64	0	0	0	2	0	0	1	0	0	0
	Hot Text	4	0.5	0.28	0.17	0.85	0	1	0	0	1	1	0	0	1	0
Mathematics																
3	Choice Multiple	2	0.24	0.03	0.21	0.26	0	0	2	0	0	0	0	0	0	0
	Choice Single	5	0.32	0.09	0.22	0.43	0	0	2	2	1	0	0	0	0	0
	Composite	1	0.28	--	0.28	0.28	0	0	1	0	0	0	0	0	0	0
	Gap Match Multiple	2	0.12	0.11	0.04	0.19	1	1	0	0	0	0	0	0	0	0

Appendix B: Summary *P* Values by Item Type

Grade	Item Type	#Items	Mean	SD	Min.	Max.	#Items by <i>P</i> Value Range									
							≤ 0.1	≤ 0.2	≤ 0.3	≤ 0.4	≤ 0.5	≤ 0.6	≤ 0.7	≤ 0.8	≤ 0.9	> 0.9
	Gap Match Single	1	0.34	--	0.34	0.34	0	0	0	1	0	0	0	0	0	0
	Graphic Gap Match	2	0.14	0.08	0.08	0.19	1	1	0	0	0	0	0	0	0	0
4	Choice Multiple	2	0.16	0.08	0.1	0.21	0	1	1	0	0	0	0	0	0	0
	Choice Single	1	0.22	--	0.22	0.22	0	0	1	0	0	0	0	0	0	0
5	Choice Multiple	2	0.54	0.14	0.44	0.63	0	0	0	0	1	0	1	0	0	0
	Choice Single	2	0.31	0.06	0.27	0.35	0	0	1	1	0	0	0	0	0	0
	Gap Match Multiple	1	0.35	--	0.35	0.35	0	0	0	1	0	0	0	0	0	0
	Hot Text	1	0.2	--	0.2	0.2	0	1	0	0	0	0	0	0	0	0
6	Choice Multiple	11	0.23	0.17	0.07	0.67	2	6	1	0	1	0	1	0	0	0
	Choice Single	13	0.43	0.23	0.14	0.74	0	1	5	2	0	0	3	2	0	0
	Composite	2	0.24	0.01	0.24	0.25	0	0	2	0	0	0	0	0	0	0
	Graphic Gap Match	1	0.08	--	0.08	0.08	1	0	0	0	0	0	0	0	0	0
	Hot Text	3	0.2	0.11	0.09	0.32	1	0	1	1	0	0	0	0	0	0
	Text Entry	2	0.16	0.07	0.11	0.21	0	1	1	0	0	0	0	0	0	0
7	Choice Multiple	1	0.35	--	0.35	0.35	0	0	0	1	0	0	0	0	0	0
	Choice Single	5	0.3	0.12	0.19	0.48	0	1	2	1	1	0	0	0	0	0
	Composite	1	0.23	--	0.23	0.23	0	0	1	0	0	0	0	0	0	0
	Hot Text	1	0.18	--	0.18	0.18	0	1	0	0	0	0	0	0	0	0
	Text Entry	2	0.09	0.07	0.04	0.14	1	1	0	0	0	0	0	0	0	0
8	Choice Multiple	2	0.18	0.07	0.13	0.23	0	1	1	0	0	0	0	0	0	0
	Composite	1	0.21	--	0.21	0.21	0	0	1	0	0	0	0	0	0	0
	Gap Match Multiple	1	0.51	--	0.51	0.51	0	0	0	0	0	1	0	0	0	0
	Hot Text	1	0.23	--	0.23	0.23	0	0	1	0	0	0	0	0	0	0
Science																
5	Choice Multiple	14	0.58	0.12	0.34	0.73	0	0	0	1	3	2	5	3	0	0
	Choice Single	44	0.52	0.14	0.17	0.79	0	1	2	3	12	13	10	3	0	0
	Composite	11	0.48	0.08	0.37	0.63	0	0	0	2	5	3	1	0	0	0
	Gap Match Multiple	17	0.5	0.21	0.02	0.91	1	0	2	2	3	3	5	0	0	1
	Gap Match Single	2	0.53	0.12	0.44	0.62	0	0	0	0	1	0	1	0	0	0
	Graphic Gap Match	12	0.6	0.25	0.1	0.96	0	1	0	1	1	4	0	2	2	1
	Hot Text	18	0.52	0.18	0.17	0.88	0	1	1	2	4	5	3	0	2	0
	Schema Set Member	1	0.37	--	0.37	0.37	0	0	0	1	0	0	0	0	0	0
8	Choice Multiple	11	0.41	0.15	0.16	0.64	0	1	1	4	1	2	2	0	0	0
	Choice Single	53	0.52	0.15	0.17	0.81	0	2	2	8	11	15	9	4	2	0
	Composite	20	0.42	0.15	0.1	0.72	1	1	2	5	4	4	2	1	0	0
	Gap Match Multiple	19	0.45	0.19	0.15	0.77	0	1	3	6	2	3	1	3	0	0
	Gap Match Single	2	0.28	0.07	0.23	0.33	0	0	1	1	0	0	0	0	0	0
	Graphic Gap Match	16	0.53	0.17	0.26	0.82	0	0	1	4	2	4	1	3	1	0
	Hot Text	13	0.55	0.14	0.24	0.73	0	0	1	1	2	5	3	1	0	0

Appendix C: Summary of Item-Total Correlations by Item Type

Table C.1. Summary of Item-Total Correlations by Item Type—Operational Items

Grade	Item Type	#Items	Mean	SD	Min.	Max.	#Items by Item-Total Correlation Range						
							≤ 0.1	≤ 0.2	≤ 0.3	≤ 0.4	≤ 0.5	≤ 0.6	> 0.6
ELA													
3	Choice Multiple	44	0.44	0.12	0.19	0.66	0	1	5	10	13	11	4
	Choice Single	519	0.37	0.10	0.05	0.61	1	21	99	206	149	40	3
	Composite	42	0.44	0.09	0.23	0.64	0	0	2	13	17	9	1
	Gap Match Multiple	31	0.40	0.09	0.23	0.60	0	0	6	8	14	2	1
	Gap Match Single	1	0.31	--	0.31	0.31	0	0	0	1	0	0	0
	Hot Text	2	0.44	0.16	0.32	0.55	0	0	0	1	0	1	0
4	Choice Multiple	63	0.43	0.09	0.17	0.63	0	1	4	21	22	13	2
	Choice Single	419	0.36	0.09	0.15	0.67	0	15	88	193	98	21	4
	Composite	34	0.46	0.07	0.23	0.57	0	0	1	5	19	9	0
	Gap Match Multiple	24	0.40	0.13	0.13	0.65	0	3	2	7	9	2	1
	Gap Match Single	1	0.39	--	0.39	0.39	0	0	0	1	0	0	0
	Hot Text	2	0.54	0.01	0.53	0.55	0	0	0	0	0	2	0
5	Choice Multiple	60	0.42	0.08	0.21	0.57	0	0	4	19	28	9	0
	Choice Single	429	0.35	0.09	0.12	0.65	0	20	102	208	70	25	4
	Composite	23	0.42	0.10	0.12	0.59	0	1	1	7	11	3	0
	Gap Match Multiple	24	0.40	0.10	0.22	0.59	0	0	4	7	9	4	0
	Gap Match Single	3	0.36	0.09	0.30	0.46	0	0	1	1	1	0	0
	Hot Text	7	0.38	0.04	0.32	0.44	0	0	0	5	2	0	0
6	Choice Multiple	51	0.41	0.08	0.24	0.58	0	0	6	17	22	6	0
	Choice Single	461	0.34	0.09	0.08	0.65	1	21	117	217	87	16	2
	Composite	43	0.44	0.09	0.21	0.61	0	0	4	10	19	9	1
	Gap Match Multiple	26	0.44	0.10	0.25	0.64	0	0	1	8	10	4	3
	Gap Match Single	1	0.38	--	0.38	0.38	0	0	0	1	0	0	0
	Hot Text	4	0.32	0.16	0.17	0.47	0	2	0	0	2	0	0
7	Choice Multiple	47	0.42	0.11	0.02	0.64	2	0	1	12	24	5	3
	Choice Single	392	0.35	0.09	0.13	0.64	0	15	103	160	95	17	2
	Composite	37	0.46	0.06	0.35	0.59	0	0	0	5	22	10	0
	Gap Match Multiple	17	0.44	0.09	0.31	0.62	0	0	0	6	9	1	1
	Gap Match Single	4	0.24	0.07	0.15	0.31	0	1	2	1	0	0	0
	Hot Text	9	0.43	0.11	0.33	0.59	0	0	0	5	1	3	0
8	Choice Multiple	52	0.39	0.09	0.22	0.64	0	0	11	17	20	3	1
	Choice Single	440	0.34	0.09	0.04	0.66	3	24	112	191	88	16	6
	Composite	47	0.43	0.07	0.24	0.57	0	0	2	15	21	9	0
	Gap Match Multiple	28	0.44	0.08	0.27	0.58	0	0	3	4	13	8	0

Appendix C: Summary Item-Total Correlations by Item Type

Grade	Item Type	#Items	Mean	SD	Min.	Max.	#Items by Item-Total Correlation Range						
							≤ 0.1	≤ 0.2	≤ 0.3	≤ 0.4	≤ 0.5	≤ 0.6	> 0.6
	Gap Match Single	3	0.35	0.06	0.31	0.41	0	0	0	2	1	0	0
	Hot Text	12	0.42	0.09	0.29	0.58	0	0	1	4	6	1	0
Mathematics													
3	Choice Multiple	46	0.29	0.15	-0.06	0.65	4	8	9	16	5	3	1
	Choice Single	534	0.27	0.12	-0.07	0.63	47	81	179	155	62	9	1
	Composite	56	0.45	0.15	0.03	1.00	1	2	5	9	19	14	6
	Gap Match Multiple	47	0.29	0.13	-0.02	0.52	5	5	16	12	7	2	0
	Gap Match Single	6	0.28	0.14	0.02	0.45	1	0	1	3	1	0	0
	Graphic Gap Match	51	0.20	0.11	-0.04	0.42	10	18	12	10	1	0	0
	Hot Text	9	0.39	0.15	0.18	0.55	0	1	2	1	1	4	0
	Text Entry	46	0.26	0.12	0.00	0.47	5	9	14	13	5	0	0
4	Choice Multiple	46	0.26	0.14	-0.02	0.55	9	8	13	9	4	3	0
	Choice Single	307	0.22	0.11	-0.08	0.50	47	71	109	71	9	0	0
	Composite	53	0.43	0.19	0.00	1.00	5	1	4	8	16	15	4
	Gap Match Multiple	31	0.32	0.12	0.09	0.58	1	3	10	9	6	2	0
	Gap Match Single	4	0.37	0.40	0.07	0.95	1	1	1	0	0	0	1
	Graphic Gap Match	43	0.27	0.12	0.01	0.52	5	7	15	11	4	1	0
	Hot Text	17	0.35	0.16	0.05	0.54	2	2	3	2	5	3	0
	Text Entry	57	0.28	0.11	-0.01	0.52	5	7	21	18	5	1	0
5	Choice Multiple	48	0.26	0.11	-0.04	0.55	5	7	16	17	2	1	0
	Choice Single	366	0.25	0.12	-0.08	0.65	43	65	129	98	23	6	2
	Composite	67	0.41	0.14	0.00	0.72	2	4	6	14	24	13	4
	Gap Match Multiple	35	0.28	0.13	0.03	0.59	4	6	11	9	3	2	0
	Gap Match Single	4	0.28	0.04	0.25	0.33	0	0	3	1	0	0	0
	Graphic Gap Match	28	0.34	0.14	0.16	0.63	0	5	8	6	3	4	2
	Hot Text	11	0.30	0.15	0.06	0.50	2	1	4	0	3	1	0
	Text Entry	52	0.28	0.10	0.02	0.47	2	9	18	19	4	0	0
6	Choice Multiple	66	0.34	0.10	0.09	0.58	1	2	20	24	16	3	0
	Choice Single	605	0.28	0.13	-1.00	0.72	49	72	214	194	63	11	2
	Composite	62	0.39	0.20	0.00	1.00	8	3	5	12	15	13	6
	Gap Match Multiple	41	0.31	0.16	0.07	0.90	2	6	11	18	1	1	2
	Gap Match Single	1	0.25	--	0.25	0.25	0	0	1	0	0	0	0
	Graphic Gap Match	29	0.33	0.09	0.14	0.53	0	2	7	13	6	1	0
	Hot Text	23	0.36	0.14	0.00	0.58	1	1	8	3	7	3	0
	Text Entry	69	0.31	0.10	0.04	0.53	1	8	21	32	6	1	0
7	Choice Multiple	40	0.37	0.13	0.06	0.56	3	2	3	15	12	5	0
	Choice Single	448	0.28	0.13	-0.89	0.77	30	46	154	167	45	5	1
	Composite	48	0.43	0.15	0.00	0.85	1	3	3	10	14	15	2

Appendix C: Summary Item-Total Correlations by Item Type

Grade	Item Type	#Items	Mean	SD	Min.	Max.	#Items by Item-Total Correlation Range						
							≤ 0.1	≤ 0.2	≤ 0.3	≤ 0.4	≤ 0.5	≤ 0.6	> 0.6
	Gap Match Multiple	36	0.30	0.17	-0.39	0.59	3	1	8	18	4	2	0
	Graphic Gap Match	9	0.28	0.16	-0.03	0.54	1	1	2	4	0	1	0
	Hot Text	22	0.33	0.13	-0.01	0.56	2	1	4	7	7	1	0
	Text Entry	74	0.29	0.12	-0.01	0.49	7	6	18	35	8	0	0
8	Choice Multiple	39	0.26	0.14	0.00	0.56	6	6	12	9	3	3	0
	Choice Single	332	0.23	0.12	-0.11	1.00	45	73	126	76	10	1	1
	Composite	55	0.38	0.17	-0.10	0.58	5	2	4	12	18	14	0
	Gap Match Multiple	42	0.24	0.13	-0.00	0.50	6	8	12	12	4	0	0
	Gap Match Single	3	0.19	0.23	-0.03	0.42	1	1	0	0	1	0	0
	Graphic Gap Match	12	0.25	0.10	0.11	0.47	0	5	3	3	1	0	0
	Hot Text	36	0.29	0.18	-0.06	1.00	3	8	8	11	4	1	1
	Text Entry	57	0.25	0.12	-0.03	0.68	6	8	24	16	2	0	1
Science													
5	Choice Multiple	8	0.48	0.11	0.28	0.62	0	0	1	0	2	4	1
	Choice Single	12	0.41	0.09	0.29	0.53	0	0	2	3	5	2	0
	Composite	4	0.53	0.09	0.42	0.63	0	0	0	0	2	1	1
	Gap Match Multiple	3	0.47	0.11	0.40	0.60	0	0	0	0	2	1	0
	Graphic Gap Match	3	0.48	0.07	0.42	0.56	0	0	0	0	2	1	0
	Hot Text	1	0.49	--	0.49	0.49	0	0	0	0	1	0	0
8	Choice Multiple	3	0.41	0.07	0.34	0.49	0	0	0	2	1	0	0
	Choice Single	15	0.43	0.07	0.29	0.55	0	0	1	3	7	4	0
	Composite	6	0.47	0.07	0.38	0.55	0	0	0	1	3	2	0
	Gap Match Multiple	3	0.39	0.01	0.38	0.40	0	0	0	2	1	0	0
	Graphic Gap Match	3	0.46	0.09	0.36	0.55	0	0	0	1	1	1	0

Table C.2. Summary of Item-Total Correlations by Item Type—Field Test Items

Grade	Item Type	#Items	Mean	SD	Min.	Max.	#Items by Item-Total Correlation Range						
							≤ 0.1	≤ 0.2	≤ 0.3	≤ 0.4	≤ 0.5	≤ 0.6	> 0.6
ELA													
3	Choice Multiple	9	0.45	0.16	0.14	0.58	0	1	1	1	2	4	0
	Choice Single	132	0.36	0.09	-0.01	0.57	3	3	26	58	36	6	0
	Composite	11	0.35	0.11	0.16	0.51	0	1	4	2	3	1	0
	Gap Match Multiple	5	0.46	0.07	0.38	0.55	0	0	0	2	2	1	0
	Hot Text	4	0.46	0.01	0.45	0.47	0	0	0	0	4	0	0
4	Choice Multiple	8	0.28	0.17	0.09	0.55	1	3	1	1	1	1	0
	Choice Single	97	0.33	0.14	-0.22	0.57	8	6	23	33	22	5	0
	Composite	11	0.47	0.12	0.29	0.67	0	0	1	2	4	2	2
	Gap Match Multiple	9	0.40	0.08	0.31	0.50	0	0	0	5	4	0	0
	Hot Text	6	0.27	0.15	0.03	0.41	1	1	1	2	1	0	0
5	Choice Multiple	11	0.41	0.12	0.15	0.54	0	1	0	4	3	3	0
	Choice Single	140	0.31	0.14	-0.06	0.55	14	9	35	44	30	8	0
	Composite	11	0.30	0.08	0.21	0.48	0	0	7	3	1	0	0
	Gap Match Multiple	4	0.47	0.09	0.38	0.58	0	0	0	1	2	1	0
	Gap Match Single	1	0.29	--	0.29	0.29	0	0	1	0	0	0	0
	Hot Text	3	0.30	0.12	0.18	0.41	0	1	0	1	1	0	0
6	Choice Multiple	17	0.38	0.10	0.21	0.60	0	0	5	6	4	1	1
	Choice Single	105	0.31	0.12	-0.09	0.55	6	9	29	39	20	2	0
	Composite	13	0.35	0.13	0.13	0.55	0	2	2	4	4	1	0
	Gap Match Multiple	3	0.35	0.04	0.32	0.40	0	0	0	3	0	0	0
	Gap Match Single	1	0.24	--	0.24	0.24	0	0	1	0	0	0	0
	Hot Text	2	0.27	0.08	0.21	0.32	0	0	1	1	0	0	0
7	Choice Multiple	21	0.46	0.15	0.19	0.70	0	1	2	4	5	6	3
	Choice Single	98	0.32	0.10	0.02	0.52	2	10	28	36	21	1	0
	Composite	10	0.46	0.12	0.26	0.66	0	0	1	2	4	1	2
	Gap Match Multiple	10	0.29	0.07	0.13	0.36	0	1	4	5	0	0	0
	Gap Match Single	2	0.32	0.22	0.17	0.48	0	1	0	0	1	0	0
	Hot Text	9	0.35	0.13	0.18	0.51	0	1	3	2	1	2	0
8	Choice Multiple	27	0.38	0.12	0.00	0.61	1	0	5	9	9	2	1
	Choice Single	141	0.31	0.14	-0.15	0.54	13	14	36	43	26	9	0
	Composite	16	0.39	0.16	0.08	0.59	1	1	4	1	4	5	0
	Gap Match Multiple	3	0.51	0.08	0.43	0.59	0	0	0	0	2	1	0
	Hot Text	4	0.28	0.17	0.06	0.44	1	0	1	1	1	0	0
Mathematics													
3	Choice Multiple	2	0.45	0.01	0.44	0.46	0	0	0	0	2	0	0
	Choice Single	5	0.24	0.04	0.20	0.29	0	0	5	0	0	0	0

Appendix C: Summary Item-Total Correlations by Item Type

Grade	Item Type	#Items	Mean	SD	Min.	Max.	#Items by Item-Total Correlation Range						
							≤ 0.1	≤ 0.2	≤ 0.3	≤ 0.4	≤ 0.5	≤ 0.6	> 0.6
	Composite	1	0.57	--	0.57	0.57	0	0	0	0	0	1	0
	Gap Match Multiple	2	0.31	0.15	0.20	0.41	0	0	1	0	1	0	0
	Gap Match Single	1	0.49	--	0.49	0.49	0	0	0	0	1	0	0
	Graphic Gap Match	2	0.35	0.07	0.31	0.40	0	0	0	1	1	0	0
4	Choice Multiple	2	0.25	0.09	0.19	0.32	0	1	0	1	0	0	0
	Choice Single	1	0.04	--	0.04	0.04	1	0	0	0	0	0	0
5	Choice Multiple	2	0.37	0.03	0.35	0.39	0	0	0	2	0	0	0
	Choice Single	2	0.20	0.08	0.14	0.26	0	1	1	0	0	0	0
	Gap Match Multiple	1	0.65	--	0.65	0.65	0	0	0	0	0	0	1
	Hot Text	1	0.59	--	0.59	0.59	0	0	0	0	0	1	0
6	Choice Multiple	11	0.37	0.09	0.22	0.49	0	0	3	5	3	0	0
	Choice Single	13	0.25	0.21	-0.24	0.48	3	3	1	1	5	0	0
	Composite	2	0.42	0.05	0.38	0.45	0	0	0	1	1	0	0
	Graphic Gap Match	1	0.32	--	0.32	0.32	0	0	0	1	0	0	0
	Hot Text	3	0.19	0.08	0.14	0.28	0	2	1	0	0	0	0
	Text Entry	2	0.48	0.03	0.46	0.50	0	0	0	0	1	1	0
7	Choice Multiple	1	0.24	--	0.24	0.24	0	0	1	0	0	0	0
	Choice Single	5	0.46	0.06	0.40	0.53	0	0	0	1	2	2	0
	Composite	1	0.30	--	0.30	0.30	0	0	0	1	0	0	0
	Hot Text	1	0.42	--	0.42	0.42	0	0	0	0	1	0	0
	Text Entry	2	0.32	0.01	0.31	0.33	0	0	0	2	0	0	0
8	Choice Multiple	2	0.10	0.27	-0.09	0.29	1	0	1	0	0	0	0
	Composite	1	0.35	--	0.35	0.35	0	0	0	1	0	0	0
	Gap Match Multiple	1	0.17	--	0.17	0.17	0	1	0	0	0	0	0
	Hot Text	1	0.24	--	0.24	0.24	0	0	1	0	0	0	0
Science													
5	Choice Multiple	14	0.40	0.11	0.16	0.52	0	2	0	1	10	1	0
	Choice Single	44	0.33	0.13	-0.09	0.50	3	3	6	16	15	1	0
	Composite	11	0.44	0.10	0.31	0.63	0	0	0	4	4	2	1
	Gap Match Multiple	17	0.40	0.18	0.03	0.64	1	3	0	1	7	3	2
	Gap Match Single	2	0.25	0.27	0.06	0.44	1	0	0	0	1	0	0
	Graphic Gap Match	12	0.38	0.13	0.15	0.59	0	2	2	1	6	1	0
	Hot Text	18	0.41	0.07	0.28	0.56	0	0	1	9	6	2	0
	Schema Set Member	1	0.30	--	0.30	0.30	0	0	1	0	0	0	0
8	Choice Multiple	11	0.39	0.12	0.18	0.58	0	1	2	3	2	3	0
	Choice Single	53	0.32	0.12	0.08	0.53	3	7	7	19	16	1	0
	Composite	20	0.45	0.11	0.21	0.64	0	0	2	3	7	7	1
	Gap Match Multiple	19	0.42	0.07	0.26	0.53	0	0	1	6	10	2	0

Appendix C: Summary Item-Total Correlations by Item Type

Grade	Item Type	#Items	Mean	SD	Min.	Max.	#Items by Item-Total Correlation Range						
							≤ 0.1	≤ 0.2	≤ 0.3	≤ 0.4	≤ 0.5	≤ 0.6	> 0.6
	Gap Match Single	2	0.45	0.00	0.45	0.46	0	0	0	0	2	0	0
	Graphic Gap Match	16	0.39	0.09	0.22	0.54	0	0	3	6	5	2	0
	Hot Text	13	0.44	0.15	0.10	0.58	0	1	1	2	2	7	0

Appendix D: Achievement Level Distributions & Scale Score Descriptive Statistics by Demographics

Table D.1. Achievement Level Distributions & Scale Score Descriptive Statistics by Demographics—ELA

ELA									
Grade	Demographic Sub-Group ^a		N	SS Descriptive Statistics		Percent of Students in Each Achievement Level ^b			
				Mean	SD	Level 3	Level 2	Level 1	L2 + L1
3	Overall		23,260	2463.42	90.77	37.7	40.5	21.8	62.3
	Gender	Female	11,416	2468.64	89.58	35.4	41.4	23.2	64.6
		Male	11,844	2458.40	91.63	39.9	39.7	20.4	60.1
	Ethnicity	AI/AN	274	2397.21	91.00	65.7	28.5	5.8	34.3
		Asian	762	2469.54	104.22	38.6	32.2	29.3	61.4
		Black	1,479	2411.67	90.43	62.5	28.7	8.7	37.5
		Hispanic	4,941	2425.68	86.75	55.1	34.7	10.2	44.9
		NH/PI	34	2422.29	97.45	52.9	35.3	11.8	47.1
		White	14,629	2482.87	84.50	28.5	44.3	27.2	71.5
		2 or more Races	1,137	2457.97	89.69	40.4	41.3	18.3	59.6
	FRL	Yes	11,054	2432.10	87.64	51.8	36.8	11.4	48.2
		No	12,202	2491.83	83.93	24.9	43.9	31.2	75.1
	LEP	Yes	4,107	2417.80	87.62	59.6	31.6	8.8	40.4
		No	19,150	2473.22	88.41	33.0	42.5	24.6	67.0
	SPED	Yes	4,381	2406.61	91.93	64.6	26.8	8.6	35.4
		No	18,879	2476.61	85.25	31.4	43.7	24.8	68.6
4	Overall		22,918	2493.26	92.39	45.0	33.0	22.0	55.0
	Gender	Female	11,157	2498.46	90.38	42.8	33.8	23.5	57.2
		Male	11,761	2488.32	94.00	47.1	32.4	20.6	52.9
	Ethnicity	AI/AN	273	2449.16	91.40	63.4	28.9	7.7	36.6
		Asian	776	2498.77	105.03	42.1	29.8	28.1	57.9
		Black	1,476	2440.54	95.63	67.7	22.8	9.5	32.3
		Hispanic	4,831	2453.82	90.85	62.4	27.7	9.9	37.6
		NH/PI	38	2460.39	91.58	57.9	31.6	10.5	42.1
		White	14,387	2512.93	85.28	36.3	36.2	27.6	63.7
		2 or more Races	1,134	2488.41	87.50	48.6	32.9	18.5	51.4
	FRL	Yes	10,837	2460.90	90.77	59.8	28.6	11.6	40.2
		No	12,078	2522.30	83.78	31.7	37.1	31.3	68.3
	LEP	Yes	3,905	2442.19	92.99	67.6	24.4	8.1	32.4
		No	19,011	2503.75	88.71	40.3	34.8	24.8	59.7
	SPED	Yes	4,085	2422.16	94.06	76.1	17.2	6.7	23.9
		No	18,833	2508.68	84.47	38.2	36.5	25.3	61.8
5	Overall		22,977	2510.98	87.99	43.2	36.2	20.7	56.8
	Gender	Female	11,136	2517.11	85.42	40.4	37.6	22.0	59.6
		Male	11,841	2505.22	89.97	45.8	34.8	19.4	54.2
	Ethnicity	AI/AN	290	2451.23	77.93	75.9	19.3	4.8	24.1
		Asian	746	2517.53	97.59	39.8	34.6	25.6	60.2
		Black	1,444	2461.54	87.51	67.0	24.6	8.4	33.0
		Hispanic	4,665	2473.80	84.71	61.4	29.5	9.1	38.6
		NH/PI	48	2485.77	98.86	60.4	20.8	18.8	39.6
		White	14,642	2529.17	82.23	34.2	40.0	25.9	65.8
		2 or more Races	1,140	2504.21	89.22	46.8	35.6	17.5	53.2
	FRL	Yes	10,558	2479.24	84.87	59.0	30.3	10.7	41.0
		No	12,415	2538.01	81.31	29.6	41.2	29.2	70.4
	LEP	Yes	3,343	2456.69	83.84	69.2	24.5	6.3	30.8
		No	19,632	2520.23	85.30	38.7	38.2	23.1	61.3
	SPED	Yes	3,897	2439.27	85.38	76.8	17.7	5.5	23.2

Appendix D: Achievement Level Distributions & Scale Score Descriptive Statistics by
Demographics

ELA									
Grade	Demographic Sub-Group ^a		N	SS Descriptive Statistics		Percent of Students in Each Achievement Level ^b			
				Mean	SD	Level 3	Level 2	Level 1	L2 + L1
		No	19,080	2525.63	81.06	36.3	39.9	23.8	63.7
6	Overall		22,851	2518.30	81.57	44.7	37.7	17.6	55.3
	Gender	Female	11,121	2525.24	78.54	41.4	39.4	19.2	58.6
		Male	11,730	2511.72	83.81	47.7	36.1	16.2	52.3
	Ethnicity	AI/AN	255	2463.19	77.02	75.3	21.2	3.5	24.7
		Asian	708	2526.42	91.31	40.7	34.2	25.1	59.3
		Black	1,423	2465.87	82.04	71.5	22.3	6.1	28.5
		Hispanic	4,687	2485.68	80.49	62.1	29.8	8.1	37.9
		NH/PI	41	2488.54	81.10	56.1	36.6	7.3	43.9
		White	14,646	2534.63	75.48	35.9	42.4	21.7	64.1
		2 or more Races	1,086	2516.43	82.19	47.1	35.9	16.9	52.9
	FRL	Yes	10,164	2488.67	80.34	60.4	30.6	9.1	39.6
		No	12,681	2542.07	74.46	32.0	43.5	24.5	68.0
	LEP	Yes	2,820	2460.47	76.40	74.6	21.8	3.5	25.4
		No	20,027	2526.45	78.94	40.4	40.0	19.6	59.6
	SPED	Yes	3,582	2448.25	79.15	79.5	16.1	4.3	20.5
		No	19,269	2531.33	75.13	38.2	41.7	20.1	61.8
7	Overall		23,430	2527.56	81.88	45.8	38.7	15.5	54.2
	Gender	Female	11,419	2534.26	79.40	42.3	41.3	16.4	57.7
		Male	12,011	2521.20	83.67	49.1	36.2	14.7	50.9
	Ethnicity	AI/AN	285	2481.76	75.85	71.9	23.5	4.6	28.1
		Asian	707	2538.19	91.23	38.8	38.3	22.9	61.2
		Black	1,559	2473.82	82.65	71.7	23.6	4.7	28.3
		Hispanic	4,863	2493.47	80.12	63.0	30.5	6.6	37.0
		NH/PI	43	2515.12	76.28	51.2	37.2	11.6	48.8
		White	14,871	2545.26	75.31	36.9	43.5	19.6	63.1
		2 or more Races	1,095	2520.90	82.93	50.7	35.7	13.6	49.3
	FRL	Yes	10,431	2498.27	80.52	61.1	31.3	7.6	38.9
		No	12,990	2551.10	75.09	33.5	44.6	21.9	66.5
	LEP	Yes	2,434	2459.94	76.31	79.3	18.7	2.0	20.7
		No	20,991	2535.42	78.83	41.9	41.0	17.1	58.1
	SPED	Yes	3,378	2453.63	77.65	82.3	14.7	3.0	17.7
		No	20,052	2540.02	75.77	39.6	42.7	17.6	60.4
8	Overall		23,886	2544.79	80.09	36.5	47.8	15.7	63.5
	Gender	Female	11,608	2553.95	76.66	31.9	50.1	18.0	68.1
		Male	12,278	2536.13	82.26	40.8	45.7	13.5	59.2
	Ethnicity	AI/AN	302	2497.30	80.54	63.2	31.1	5.6	36.8
		Asian	663	2550.01	86.21	33.8	48.0	18.3	66.2
		Black	1,542	2497.94	79.56	61.7	32.8	5.5	38.3
		Hispanic	5,149	2513.37	79.43	52.5	40.4	7.2	47.5
		NH/PI	35	2519.17	87.52	51.4	37.1	11.4	48.6
		White	15,141	2561.41	74.43	27.9	52.4	19.8	72.1
		2 or more Races	1,050	2539.21	78.93	39.5	46.1	14.4	60.5
	FRL	Yes	10,397	2515.81	79.37	51.8	40.1	8.1	48.2
		No	13,485	2567.13	73.19	24.7	53.8	21.5	75.3
	LEP	Yes	2,203	2471.76	74.85	74.8	24.0	1.2	25.2
		No	21,680	2552.21	76.81	32.6	50.2	17.1	67.4
	SPED	Yes	3,285	2470.68	75.42	76.9	19.9	3.2	23.1
		No	20,601	2556.61	74.26	30.1	52.3	17.7	69.9

^a AI/AN = American Indian or Alaska Native; NH/PI = Native Hawaiian or Other Pacific Islander; FRL = free and reduced lunch; LEP = limited English proficient; SPED = special education

^b Level 3 = *Developing*; Level 2 = *On Track*; Level 1 = *Advanced*

Appendix D: Achievement Level Distributions & Scale Score Descriptive Statistics by
Demographics

**Table D.2. Achievement Level Distributions & Scale Score Descriptive Statistics by
Demographics—Mathematics**

Mathematics									
Grade	Demographic Sub-Group ^a		N	SS Descriptive Statistics		Percent of Students in Each Achievement Level ^b			
				Mean	SD	Level 3	Level 2	Level 1	L2 + L1
3	Overall		23,197	1193.73	88.26	41.8	45.1	13.1	58.2
	Gender	Female	11,381	1185.38	83.76	45.5	44.0	10.4	54.5
		Male	11,816	1201.78	91.67	38.2	46.0	15.8	61.8
	Ethnicity	AI/AN	274	1121.32	79.52	75.5	21.5	2.9	24.5
		Asian	761	1204.94	102.66	38.5	41.8	19.7	61.5
		Black	1,478	1133.95	80.10	70.2	27.3	2.6	29.8
		Hispanic	4,870	1154.39	77.29	62.1	33.7	4.3	37.9
		NH/PI	34	1159.65	93.02	58.8	29.4	11.8	41.2
		White	14,638	1214.79	83.58	31.2	51.5	17.2	68.8
		2 or more Races	1,138	1180.27	87.09	47.8	41.9	10.3	52.2
	FRL	Yes	11,007	1160.98	80.86	58.0	36.4	5.6	42.0
		No	12,186	1223.35	84.06	27.1	52.9	20.0	72.9
	LEP	Yes	4,035	1149.35	81.40	64.5	30.8	4.7	35.5
		No	19,159	1203.09	86.77	37.0	48.1	14.9	63.0
	SPED	Yes	4,381	1139.23	88.46	68.2	26.4	5.4	31.8
		No	18,816	1206.42	83.24	35.6	49.4	15.0	64.4
4	Overall		22,842	1224.02	87.18	42.0	46.0	12.1	58.0
	Gender	Female	11,119	1217.01	82.59	44.2	46.7	9.1	55.8
		Male	11,723	1230.68	90.82	39.8	45.3	14.8	60.2
	Ethnicity	AI/AN	275	1171.40	78.30	67.6	30.5	1.8	32.4
		Asian	776	1238.50	101.78	38.0	43.2	18.8	62.0
		Black	1,475	1158.20	78.68	74.1	22.6	3.3	25.9
		Hispanic	4,762	1186.12	78.16	61.0	34.7	4.3	39.0
		NH/PI	38	1195.18	79.48	55.3	42.1	2.6	44.7
		White	14,382	1244.82	82.18	31.3	52.9	15.7	68.7
		2 or more Races	1,132	1208.93	82.24	51.0	41.3	7.8	49.0
	FRL	Yes	10,791	1191.01	81.09	58.3	36.4	5.3	41.7
		No	12,049	1253.60	81.65	27.3	54.5	18.1	72.7
	LEP	Yes	3,834	1178.14	80.57	64.8	31.1	4.1	35.2
		No	19,006	1233.29	85.52	37.3	49.0	13.7	62.7
	SPED	Yes	4,086	1163.94	82.88	72.8	23.2	4.0	27.2
		No	18,756	1237.11	82.47	35.2	50.9	13.8	64.8
5	Overall		22,917	1242.12	83.43	34.6	46.5	18.8	65.4
	Gender	Female	11,107	1237.32	78.57	35.5	49.1	15.4	64.5
		Male	11,810	1246.64	87.52	33.8	44.1	22.1	66.2
	Ethnicity	AI/AN	291	1175.21	70.23	70.8	26.8	2.4	29.2
		Asian	745	1257.54	99.63	31.4	40.5	28.1	68.6
		Black	1,442	1183.69	76.37	65.3	28.9	5.8	34.7
		Hispanic	4,608	1206.15	75.95	52.9	39.4	7.7	47.1
		NH/PI	48	1240.33	98.62	39.6	39.6	20.8	60.4
		White	14,643	1260.82	78.25	24.8	51.4	23.8	75.2
		2 or more Races	1,139	1228.43	83.13	41.2	44.0	14.8	58.8
	FRL	Yes	10,531	1209.87	76.94	50.5	40.7	8.8	49.5
		No	12,383	1269.59	78.74	21.1	51.4	27.4	78.9
	LEP	Yes	3,285	1193.94	75.99	59.5	34.2	6.3	40.5
		No	19,631	1250.19	81.88	30.5	48.6	21.0	69.5
	SPED	Yes	3,895	1178.24	77.96	68.2	26.2	5.5	31.8
		No	19,022	1255.20	78.33	27.8	50.7	21.6	72.2
6	Overall		22,774	1242.68	85.48	42.3	40.0	17.6	57.7
	Gender	Female	11,085	1241.07	82.04	43.1	40.4	16.5	56.9

Appendix D: Achievement Level Distributions & Scale Score Descriptive Statistics by
Demographics

Mathematics									
Grade	Demographic Sub-Group ^a		N	SS Descriptive Statistics		Percent of Students in Each Achievement Level ^b			
				Mean	SD	Level 3	Level 2	Level 1	L2 + L1
		Male	11,689	1244.21	88.59	41.6	39.7	18.7	58.4
	Ethnicity	AI/AN	254	1177.80	76.58	74.8	20.9	4.3	25.2
		Asian	707	1252.38	107.16	42.9	31.4	25.7	57.1
		Black	1,416	1171.75	79.10	77.3	18.7	4.0	22.7
		Hispanic	4,629	1207.42	79.20	61.3	30.6	8.1	38.7
		NH/PI	41	1222.49	88.92	46.3	43.9	9.8	53.7
		White	14,639	1262.52	78.92	31.7	46.1	22.2	68.3
		2 or more Races	1,085	1227.57	82.47	51.9	35.9	12.3	48.1
	FRL	Yes	10,130	1210.27	80.43	58.8	32.8	8.5	41.2
		No	12,641	1268.66	80.39	29.2	45.8	25.0	70.8
	LEP	Yes	2,763	1184.12	75.79	72.9	22.9	4.3	27.1
		No	20,008	1250.77	83.57	38.1	42.4	19.5	61.9
	SPED	Yes	3,574	1171.78	79.19	77.5	18.7	3.8	22.5
		No	19,200	1255.88	79.94	35.8	44.0	20.2	64.2
7	Overall		23,348	1246.44	83.63	34.0	46.0	20.0	66.0
	Gender	Female	11,383	1242.75	79.84	35.2	47.1	17.6	64.8
		Male	11,965	1249.94	86.94	32.8	44.9	22.3	67.2
	Ethnicity	AI/AN	286	1196.82	70.13	60.8	35.0	4.2	39.2
		Asian	705	1268.52	106.74	32.5	37.0	30.5	67.5
		Black	1,562	1179.41	73.28	69.0	26.6	4.4	31.0
		Hispanic	4,805	1211.29	74.21	51.5	39.6	8.9	48.5
		NH/PI	42	1230.43	72.89	40.5	42.9	16.7	59.5
		White	14,852	1266.10	78.33	23.5	51.1	25.5	76.5
		2 or more Races	1,089	1229.06	83.42	43.4	41.7	14.9	56.6
	FRL	Yes	10,386	1213.92	75.29	49.4	41.3	9.3	50.6
		No	12,955	1272.53	80.80	21.6	49.7	28.7	78.4
	LEP	Yes	2,377	1183.65	66.49	67.3	29.6	3.1	32.7
		No	20,966	1253.56	82.40	30.2	47.8	22.0	69.8
	SPED	Yes	3,371	1178.00	70.45	70.1	26.3	3.6	29.9
		No	19,977	1257.98	80.08	27.9	49.3	22.8	72.1
8	Overall		23,787	1254.57	88.19	38.1	38.7	23.2	61.9
	Gender	Female	11,587	1255.41	84.10	37.4	39.9	22.6	62.6
		Male	12,200	1253.78	91.90	38.7	37.6	23.6	61.3
	Ethnicity	AI/AN	299	1192.66	80.94	68.9	23.7	7.4	31.1
		Asian	661	1272.37	109.13	33.7	31.8	34.5	66.3
		Black	1,539	1188.93	80.57	70.7	22.2	7.1	29.3
		Hispanic	5,086	1218.51	79.60	55.9	33.2	10.9	44.1
		NH/PI	33	1247.30	88.68	48.5	30.3	21.2	51.5
		White	15,114	1275.23	82.59	27.7	43.0	29.2	72.3
		2 or more Races	1,050	1234.74	87.97	47.2	36.4	16.4	52.8
	FRL	Yes	10,358	1219.36	81.53	55.1	33.5	11.4	44.9
		No	13,424	1281.75	83.39	25.0	42.8	32.2	75.0
	LEP	Yes	2,135	1184.62	73.71	73.4	22.3	4.3	26.6
		No	21,647	1261.47	86.47	34.6	40.4	25.0	65.4
	SPED	Yes	3,279	1176.74	76.80	76.6	18.8	4.6	23.4
		No	20,508	1267.02	83.39	31.9	41.9	26.1	68.1

^a AI/AN = American Indian or Alaska Native; NH/PI = Native Hawaiian or Other Pacific Islander; FRL = free and reduced lunch; LEP = limited English proficient; SPED = special education

^b Level 3 = *Developing*; Level 2 = *On Track*; Level 1 = *Advanced*

Appendix D: Achievement Level Distributions & Scale Score Descriptive Statistics by
Demographics

**Table D.3. Achievement Level Distributions & Scale Score Descriptive Statistics by Demographics—
Science**

Science									
Grade	Demographic Sub-Group ^a		N	SS Descriptive Statistics		Percent of Students in Each Achievement Level ^b			
				Mean	SD	Level 3	Level 2	Level 1	L2 + L1
5	Overall		22,888	3119.63	27.89	22.9	61.1	16.0	77.1
	Gender	Female	11,094	3118.41	26.46	22.8	63.3	13.9	77.2
		Male	11,794	3120.78	29.14	23.1	59.0	17.9	76.9
	Ethnicity	AI/AN	286	3098.82	22.95	50.7	46.5	2.8	49.3
		Asian	746	3120.92	29.74	22.9	58.6	18.5	77.1
		Black	1,443	3101.47	25.30	48.5	47.0	4.5	51.5
		Hispanic	4,606	3107.64	24.21	36.7	57.6	5.7	63.3
		NH/PI	47	3113.17	28.30	34.0	51.1	14.9	66.0
		White	14,618	3125.97	26.93	15.1	64.1	20.8	84.9
		2 or more Races	1,139	3114.45	27.30	27.7	60.2	12.0	72.3
	FRL	Yes	10,511	3109.50	25.50	34.5	58.0	7.5	65.5
		No	12,370	3128.25	26.92	13.1	63.8	23.1	86.9
	LEP	Yes	3,277	3103.10	23.30	44.0	52.5	3.5	56.0
		No	19,609	3122.40	27.64	19.4	62.5	18.1	80.6
	SPED	Yes	3,893	3100.10	25.78	52.4	42.8	4.8	47.6
		No	18,995	3123.63	26.59	16.9	64.8	18.3	83.1
8	Overall		23,807	3111.58	30.21	34.9	56.6	8.5	65.1
	Gender	Female	11,586	3111.52	28.83	34.1	58.4	7.4	65.9
		Male	12,221	3111.64	31.46	35.6	54.9	9.5	64.4
	Ethnicity	AI/AN	300	3091.51	27.03	64.7	34.3	1.0	35.3
		Asian	663	3112.64	33.43	35.0	54.0	11.0	65.0
		Black	1,543	3088.85	26.11	68.0	31.0	1.0	32.0
		Hispanic	5,084	3098.55	27.28	52.7	44.6	2.8	47.3
		NH/PI	34	3103.09	31.23	44.1	50.0	5.9	55.9
		White	15,129	3118.95	28.58	24.5	64.1	11.4	75.5
		2 or more Races	1,048	3107.24	29.51	40.5	53.4	6.1	59.5
	FRL	Yes	10,358	3100.34	28.31	50.2	46.3	3.5	49.8
		No	13,442	3120.26	28.72	23.1	64.6	12.3	76.9
	LEP	Yes	2,138	3085.10	23.66	74.0	25.6	0.4	26.0
		No	21,665	3114.20	29.52	31.0	59.7	9.3	69.0
	SPED	Yes	3,285	3086.76	26.16	72.2	26.0	1.8	27.8
		No	20,522	3115.56	28.89	28.9	61.5	9.6	71.1

^a AI/AN = American Indian or Alaska Native; NH/PI = Native Hawaiian or Other Pacific Islander; FRL = free and reduced lunch; LEP = limited English proficient; SPED = special education

^b Level 3 = *Developing*; Level 2 = *On Track*; Level 1 = *Advanced*

Appendix E: Marginal Reliability by Demographics

Table E.1. Marginal Reliability by Demographics—ELA

ELA						
Grade	Demographic Sub-Group ^a		N	Variance	MSE	Marginal Reliability
3	Overall		23,260	8239.3	758.2	0.91
	Gender	Female	11,416	8024.2	755.7	0.91
		Male	11,844	8395.9	760.5	0.91
	Ethnicity	AI/AN	274	8281.3	816.3	0.90
		Asian	762	10862.5	792.9	0.93
		Black	1,479	8178.0	788.3	0.90
		Hispanic	4,941	7525.2	766.6	0.90
		NH/PI	34	9496.4	783.3	0.92
		White	14,629	7140.9	749.2	0.90
		2 or more Races	1,137	8044.3	756.7	0.91
		FRL	Yes	11,054	7681.0	762.6
	No		12,202	7044.9	753.9	0.89
	LEP	Yes	4,107	7678.0	778.1	0.90
		No	19,150	7815.6	753.8	0.90
SPED	Yes	4,381	8451.9	797.1	0.91	
	No	18,879	7267.7	749.1	0.90	
4	Overall		22,918	8536.6	786.2	0.91
	Gender	Female	11,157	8169.2	786.3	0.90
		Male	11,761	8835.8	786.1	0.91
	Ethnicity	AI/AN	273	8353.2	766.0	0.91
		Asian	776	11031.8	822.5	0.93
		Black	1,476	9145.4	779.8	0.91
		Hispanic	4,831	8254.2	767.4	0.91
		NH/PI	38	8386.8	750.0	0.91
		White	14,387	7271.9	792.6	0.89
		2 or more Races	1,134	7655.5	774.7	0.90
		FRL	Yes	10,837	8238.4	768.0
	No		12,078	7019.5	802.5	0.89
	LEP	Yes	3,905	8647.4	774.5	0.91
		No	19,011	7869.4	788.6	0.90
SPED	Yes	4,085	8847.3	791.5	0.91	
	No	18,833	7135.3	785.1	0.89	
5	Overall		22,977	7743.0	748.2	0.90
	Gender	Female	11,136	7296.1	749.1	0.90
		Male	11,841	8095.4	747.3	0.91
	Ethnicity	AI/AN	290	6073.3	731.3	0.88
		Asian	746	9523.8	775.7	0.92
		Black	1,444	7658.5	732.6	0.90
		Hispanic	4,665	7175.0	727.8	0.90
		NH/PI	48	9773.2	776.6	0.92
		White	14,642	6762.4	755.2	0.89
		2 or more Races	1,140	7960.2	746.4	0.91
		FRL	Yes	10,558	7202.2	731.1
	No		12,415	6610.7	762.7	0.88
	LEP	Yes	3,343	7029.3	732.8	0.90
		No	19,632	7276.8	750.8	0.90
SPED	Yes	3,897	7289.9	750.7	0.90	
	No	19,080	6570.8	747.7	0.89	
6	Overall		22,851	6653.6	692.3	0.90
	Gender	Female	11,121	6168.9	688.2	0.89
		Male	11,730	7024.7	696.3	0.90
	Ethnicity	AI/AN	255	5932.4	723.2	0.88

Appendix F: Scatterplots for Scale Score CSEM

ELA						
Grade	Demographic Sub-Group ^a		N	Variance	MSE	Marginal Reliability
		Asian	708	8337.9	707.5	0.92
		Black	1,423	6730.6	720.5	0.89
		Hispanic	4,687	6478.2	699.9	0.89
		NH/PI	41	6577.1	697.9	0.89
		White	14,646	5696.7	685.6	0.88
		2 or more Races	1,086	6755.5	695.0	0.90
	FRL	Yes	10,164	6455.2	700.0	0.89
		No	12,681	5544.6	686.1	0.88
	LEP	Yes	2,820	5836.6	720.0	0.88
		No	20,027	6232.0	688.4	0.89
	SPED	Yes	3,582	6263.9	745.7	0.88
		No	19,269	5644.4	682.4	0.88
7	Overall		23,430	6703.6	708.0	0.89
	Gender	Female	11,419	6304.0	705.1	0.89
		Male	12,011	7000.8	710.8	0.90
	Ethnicity	AI/AN	285	5752.6	720.2	0.87
		Asian	707	8323.1	724.5	0.91
		Black	1,559	6831.0	731.3	0.89
		Hispanic	4,863	6418.8	714.7	0.89
		NH/PI	43	5818.8	701.2	0.88
		White	14,871	5672.3	702.3	0.88
		2 or more Races	1,095	6877.2	708.7	0.90
	FRL	Yes	10,431	6484.3	713.2	0.89
		No	12,990	5638.3	703.7	0.88
	LEP	Yes	2,434	5822.7	748.1	0.87
		No	20,991	6214.4	703.3	0.89
	SPED	Yes	3,378	6030.2	756.1	0.87
		No	20,052	5741.4	699.9	0.88
8	Overall		23,886	6413.9	682.0	0.89
	Gender	Female	11,608	5877.4	680.8	0.88
		Male	12,278	6767.4	683.2	0.90
	Ethnicity	AI/AN	302	6486.0	695.3	0.89
		Asian	663	7431.5	693.8	0.91
		Black	1,542	6329.9	682.8	0.89
		Hispanic	5,149	6309.1	678.3	0.89
		NH/PI	35	7659.6	704.0	0.91
		White	15,141	5539.5	683.0	0.88
		2 or more Races	1,050	6230.5	673.8	0.89
	FRL	Yes	10,397	6299.2	678.9	0.89
		No	13,485	5356.3	684.4	0.87
	LEP	Yes	2,203	5602.8	701.0	0.87
		No	21,680	5900.1	680.1	0.88
	SPED	Yes	3,285	5687.5	701.9	0.88
		No	20,601	5514.5	678.9	0.88

^a AI/AN = American Indian or Alaska Native; NH/PI = Native Hawaiian or Other Pacific Islander; FRL = free and reduced lunch; LEP = limited English proficient; SPED = special education

Table E.2. Marginal Reliability by Demographics—Mathematics

Mathematics						
Grade	Demographic Sub-Group ^a		N	Variance	MSE	Marginal Reliability
3	Overall		23,197	7789.5	386.3	0.95
	Gender	Female	11,381	7015.9	381.2	0.95
		Male	11,816	8403.4	391.3	0.95
	Ethnicity	AI/AN	274	6324.1	389.2	0.94
		Asian	761	10538.6	405.7	0.96
		Black	1,478	6416.0	383.7	0.94
		Hispanic	4,870	5974.1	378.6	0.94
		NH/PI	34	8652.5	388.2	0.96
		White	14,638	6986.1	388.3	0.94
		2 or more Races	1,138	7584.5	384.1	0.95
	FRL	Yes	11,007	6538.0	379.9	0.94
		No	12,186	7065.8	392.1	0.94
	LEP	Yes	4,035	6625.2	382.2	0.94
		No	19,159	7529.0	387.2	0.95
	SPED	Yes	4,381	7824.6	389.9	0.95
		No	18,816	6929.0	385.5	0.94
4	Overall		22,842	7599.9	374.7	0.95
	Gender	Female	11,119	6821.8	371.8	0.95
		Male	11,723	8247.5	377.4	0.95
	Ethnicity	AI/AN	275	6131.2	379.1	0.94
		Asian	776	10359.4	388.0	0.96
		Black	1,475	6190.2	383.4	0.94
		Hispanic	4,762	6108.5	373.6	0.94
		NH/PI	38	6316.4	372.6	0.94
		White	14,382	6752.8	373.7	0.94
		2 or more Races	1,132	6763.0	370.6	0.95
	FRL	Yes	10,791	6575.8	374.8	0.94
		No	12,049	6667.2	374.6	0.94
	LEP	Yes	3,834	6492.2	377.2	0.94
		No	19,006	7313.6	374.2	0.95
	SPED	Yes	4,086	6868.3	384.3	0.94
		No	18,756	6801.6	372.6	0.95
5	Overall		22,917	6960.5	366.2	0.95
	Gender	Female	11,107	6173.5	363.0	0.94
		Male	11,810	7659.1	369.2	0.95
	Ethnicity	AI/AN	291	4932.0	369.4	0.93
		Asian	745	9925.7	389.6	0.96
		Black	1,442	5832.4	371.8	0.94
		Hispanic	4,608	5767.8	364.1	0.94
		NH/PI	48	9725.8	406.5	0.96
		White	14,643	6123.7	364.9	0.94
		2 or more Races	1,139	6909.8	366.7	0.95
	FRL	Yes	10,531	5920.5	364.5	0.94
		No	12,383	6199.8	367.6	0.94
	LEP	Yes	3,285	5774.9	368.0	0.94
		No	19,631	6704.4	365.9	0.95
	SPED	Yes	3,895	6077.1	373.2	0.94
		No	19,022	6135.0	364.8	0.94
6	Overall		22,774	7306.6	364.9	0.95
	Gender	Female	11,085	6731.2	363.3	0.95
		Male	11,689	7848.1	366.5	0.95
	Ethnicity	AI/AN	254	5864.9	383.9	0.93
		Asian	707	11482.2	374.0	0.97
		Black	1,416	6257.5	385.5	0.94

Appendix F: Scatterplots for Scale Score CSEM

Mathematics							
Grade	Demographic Sub-Group ^a		N	Variance	MSE	Marginal Reliability	
		Hispanic	4,629	6273.3	372.0	0.94	
		NH/PI	41	7906.6	363.1	0.95	
		White	14,639	6228.7	359.9	0.94	
		2 or more Races	1,085	6801.2	365.9	0.95	
	FRL	Yes	10,130	6469.6	371.5	0.94	
		No	12,641	6462.3	359.6	0.94	
	LEP	Yes	2,763	5744.0	380.2	0.93	
		No	20,008	6984.5	362.8	0.95	
	SPED	Yes	3,574	6271.3	388.3	0.94	
		No	19,200	6389.7	360.6	0.94	
	7	Overall		23,348	6993.7	373.7	0.95
		Gender	Female	11,383	6373.7	372.4	0.94
Male			11,965	7559.0	375.0	0.95	
Ethnicity		AI/AN	286	4918.8	392.9	0.92	
		Asian	705	11393.1	378.6	0.97	
		Black	1,562	5369.8	406.4	0.92	
		Hispanic	4,805	5506.5	385.2	0.93	
		NH/PI	42	5312.4	372.0	0.93	
		White	14,852	6135.1	365.4	0.94	
		2 or more Races	1,089	6959.2	382.2	0.95	
FRL		Yes	10,386	5668.0	384.4	0.93	
		No	12,955	6528.9	365.2	0.94	
LEP		Yes	2,377	4420.3	399.3	0.91	
		No	20,966	6789.0	370.8	0.95	
SPED		Yes	3,371	4963.0	408.0	0.92	
		No	19,977	6412.9	367.9	0.94	
8	Overall		23,787	7776.7	389.7	0.95	
	Gender	Female	11,587	7072.2	386.2	0.95	
		Male	12,200	8445.1	393.0	0.95	
	Ethnicity	AI/AN	299	6552.0	431.1	0.93	
		Asian	661	11909.7	399.0	0.97	
		Black	1,539	6490.7	435.0	0.93	
		Hispanic	5,086	6336.0	408.7	0.94	
		NH/PI	33	7864.4	392.6	0.95	
		White	15,114	6821.1	376.6	0.94	
		2 or more Races	1,050	7739.4	401.5	0.95	
	FRL	Yes	10,358	6647.1	409.2	0.94	
		No	13,424	6953.7	374.7	0.95	
	LEP	Yes	2,135	5433.8	436.0	0.92	
		No	21,647	7477.9	385.1	0.95	
	SPED	Yes	3,279	5898.3	446.5	0.92	
		No	20,508	6953.9	380.6	0.95	

^a AI/AN = American Indian or Alaska Native; NH/PI = Native Hawaiian or Other Pacific Islander; FRL = free and reduced lunch; LEP = limited English proficient; SPED = special education

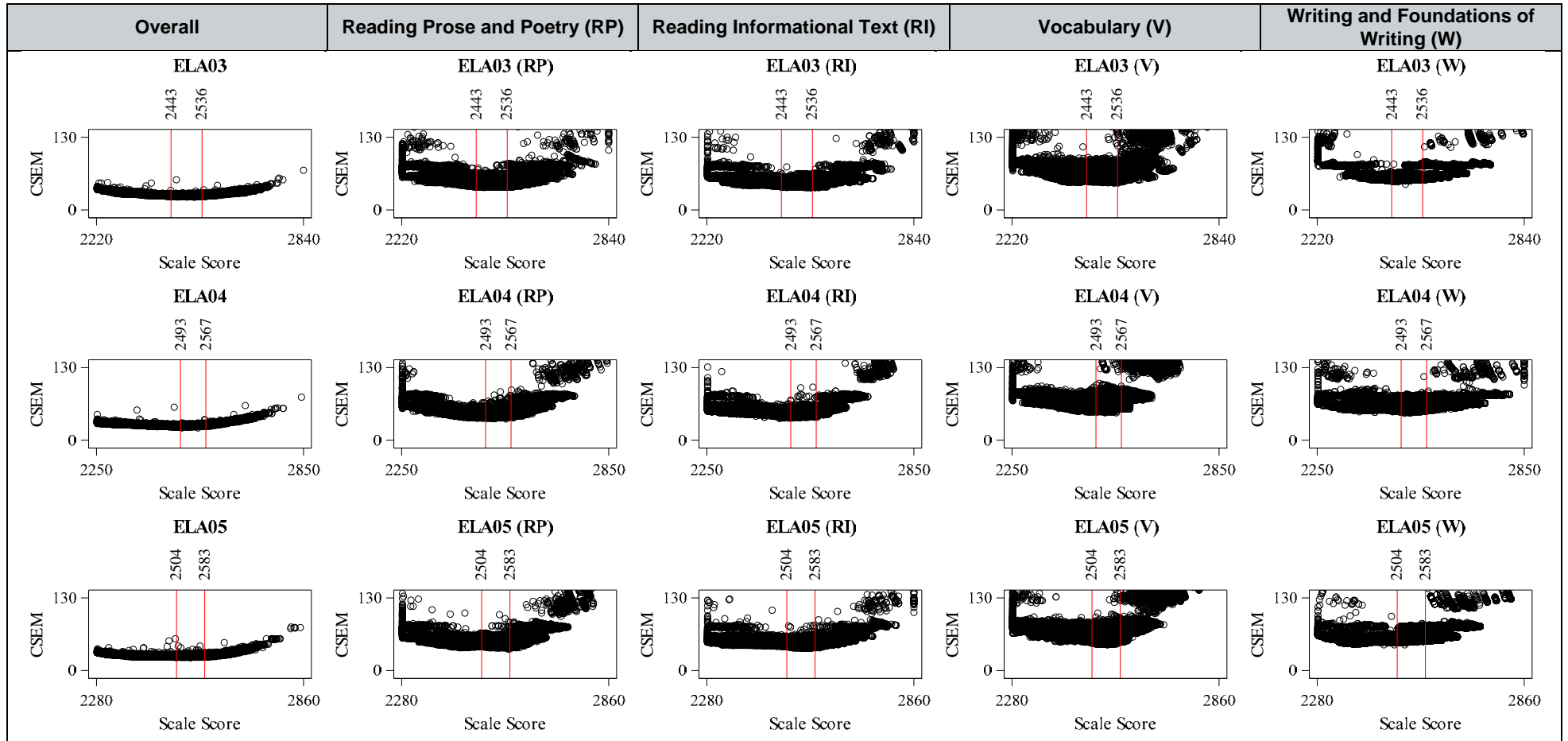
Table E.3. Marginal Reliability by Demographics—Science

Mathematics						
Grade	Demographic Sub-Group ^a		N	Variance	MSE	Marginal Reliability
5	Overall		22,888	778.1	98.1	0.87
	Gender	Female	11,094	700.0	93.4	0.87
		Male	11,794	848.9	102.6	0.88
	Ethnicity	AI/AN	286	526.8	83.1	0.84
		Asian	746	884.5	102.6	0.88
		Black	1,443	639.8	87.0	0.86
		Hispanic	4,606	585.9	83.0	0.86
		NH/PI	47	800.7	90.1	0.89
		White	14,618	725.5	104.6	0.86
		2 or more Races	1,139	745.1	91.4	0.88
		FRL	Yes	10,511	650.5	86.3
	No		12,370	724.8	108.2	0.85
	LEP	Yes	3,277	542.9	81.8	0.85
		No	19,609	763.9	100.8	0.87
SPED	Yes	3,893	664.8	88.1	0.87	
	No	18,995	707.1	100.2	0.86	
8	Overall		23,807	912.5	139.0	0.85
	Gender	Female	11,586	831.0	135.4	0.84
		Male	12,221	989.7	142.3	0.86
	Ethnicity	AI/AN	300	730.6	126.2	0.83
		Asian	663	1117.4	149.3	0.87
		Black	1,543	681.9	126.6	0.81
		Hispanic	5,084	744.1	125.5	0.83
		NH/PI	34	975.5	136.4	0.86
		White	15,129	816.6	145.0	0.82
		2 or more Races	1,048	870.9	133.0	0.85
		FRL	Yes	10,358	801.5	127.9
	No		13,442	825.1	147.5	0.82
	LEP	Yes	2,138	560.0	125.1	0.78
		No	21,665	871.2	140.3	0.84
SPED	Yes	3,285	684.2	128.3	0.81	
	No	20,522	834.6	140.7	0.83	

^a AI/AN = American Indian or Alaska Native; NH/PI = Native Hawaiian or Other Pacific Islander; FRL = free and reduced lunch; LEP = limited English proficient; SPED = special education

Appendix F: Scatterplots for Scale Score CSEM

Figure F.1. Scatterplots for Scale Score CSEM—ELA



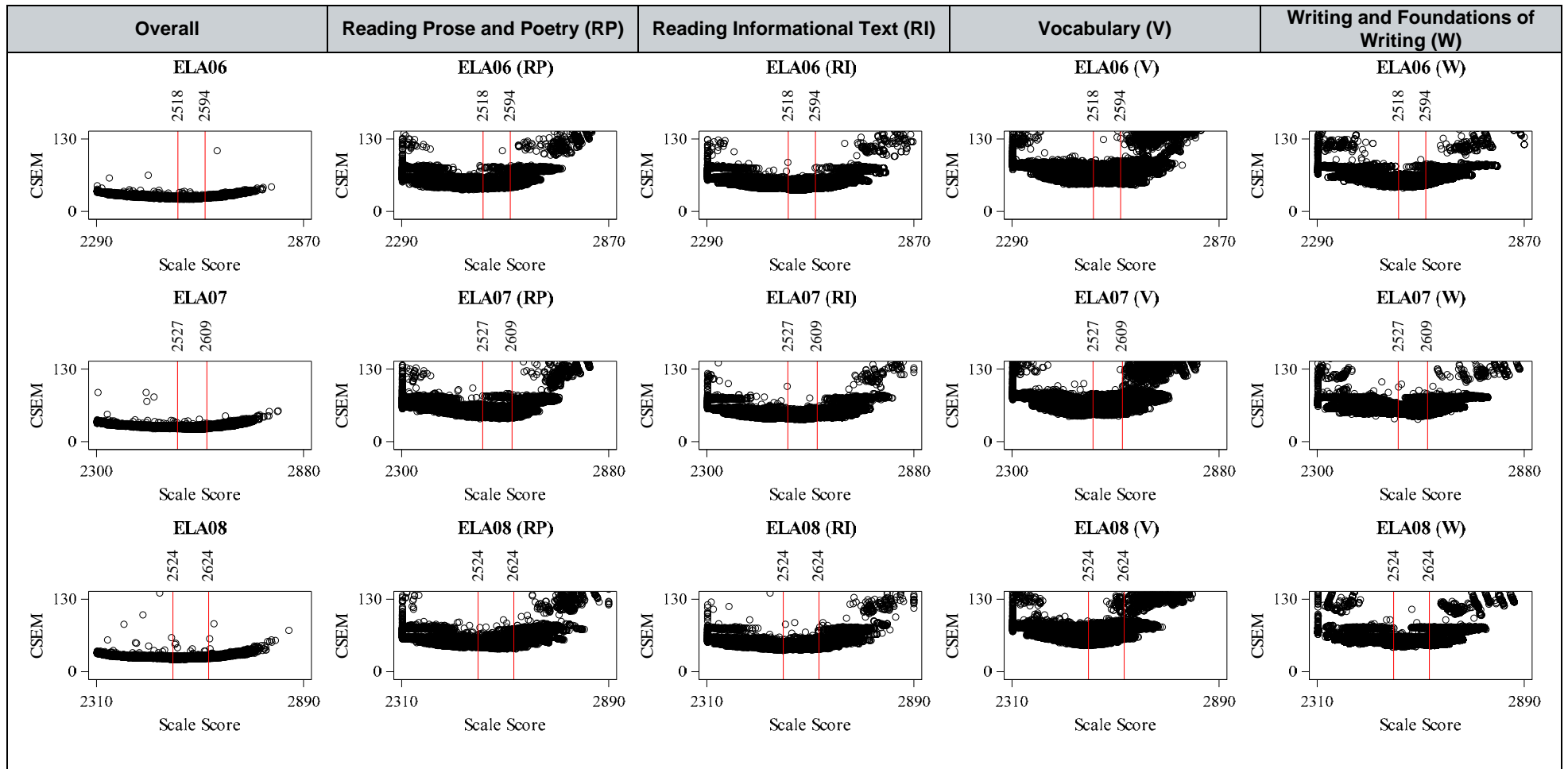
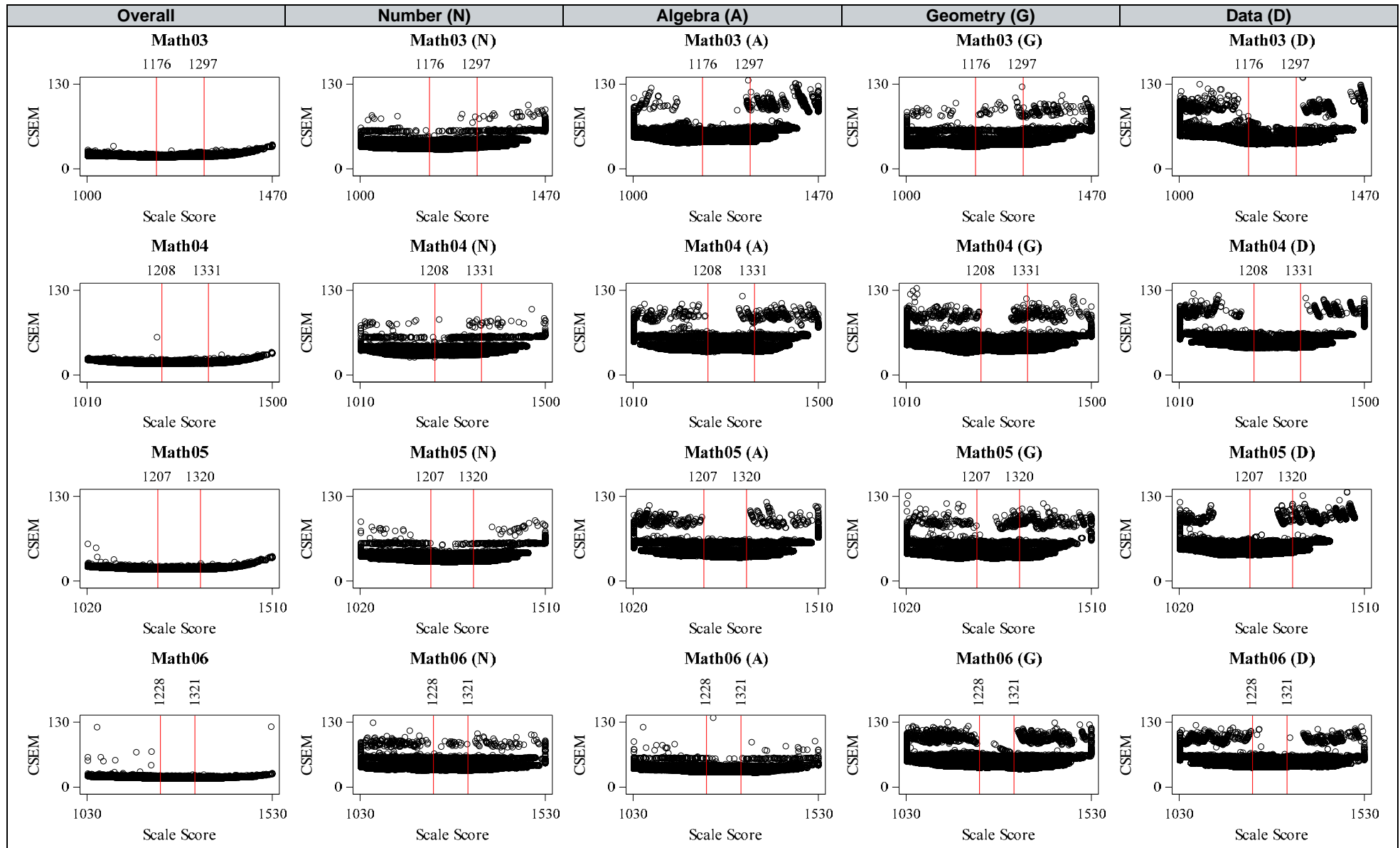


Figure F.2. Scatterplots for Scale Score CSEM—Mathematics



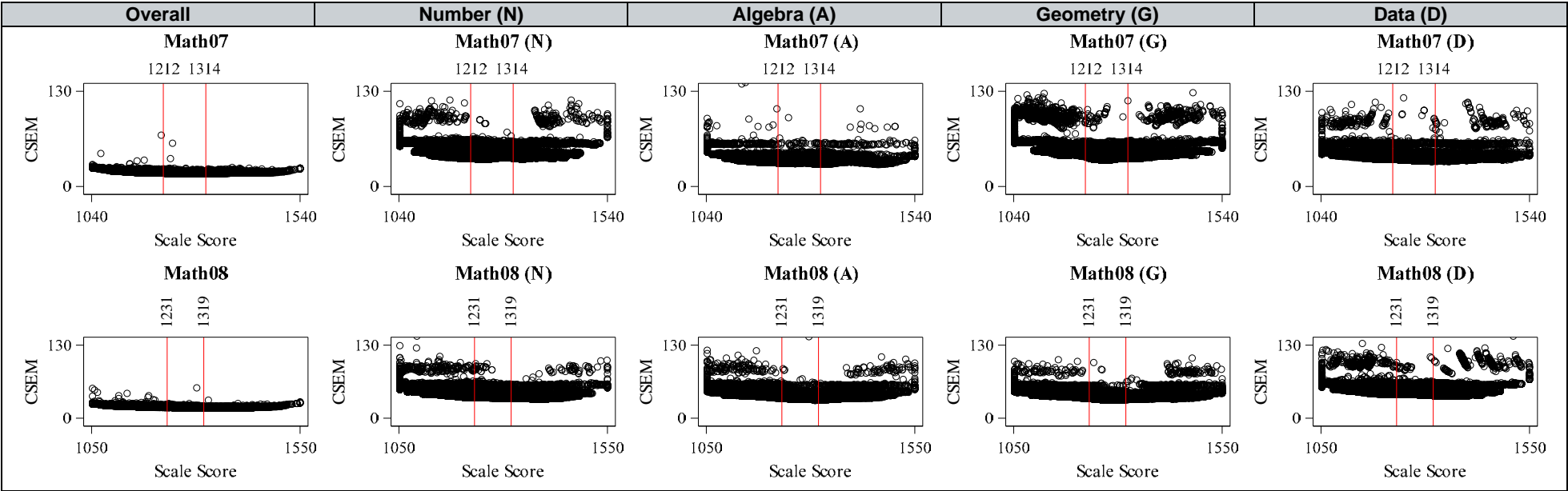
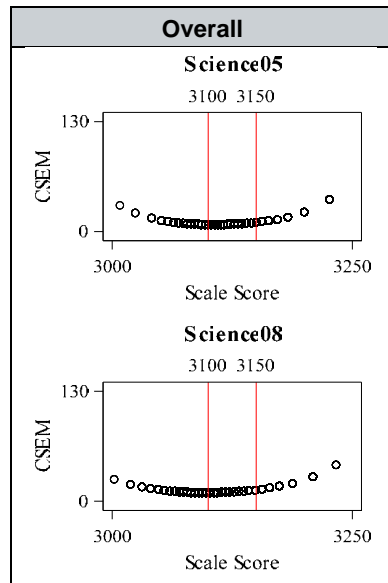
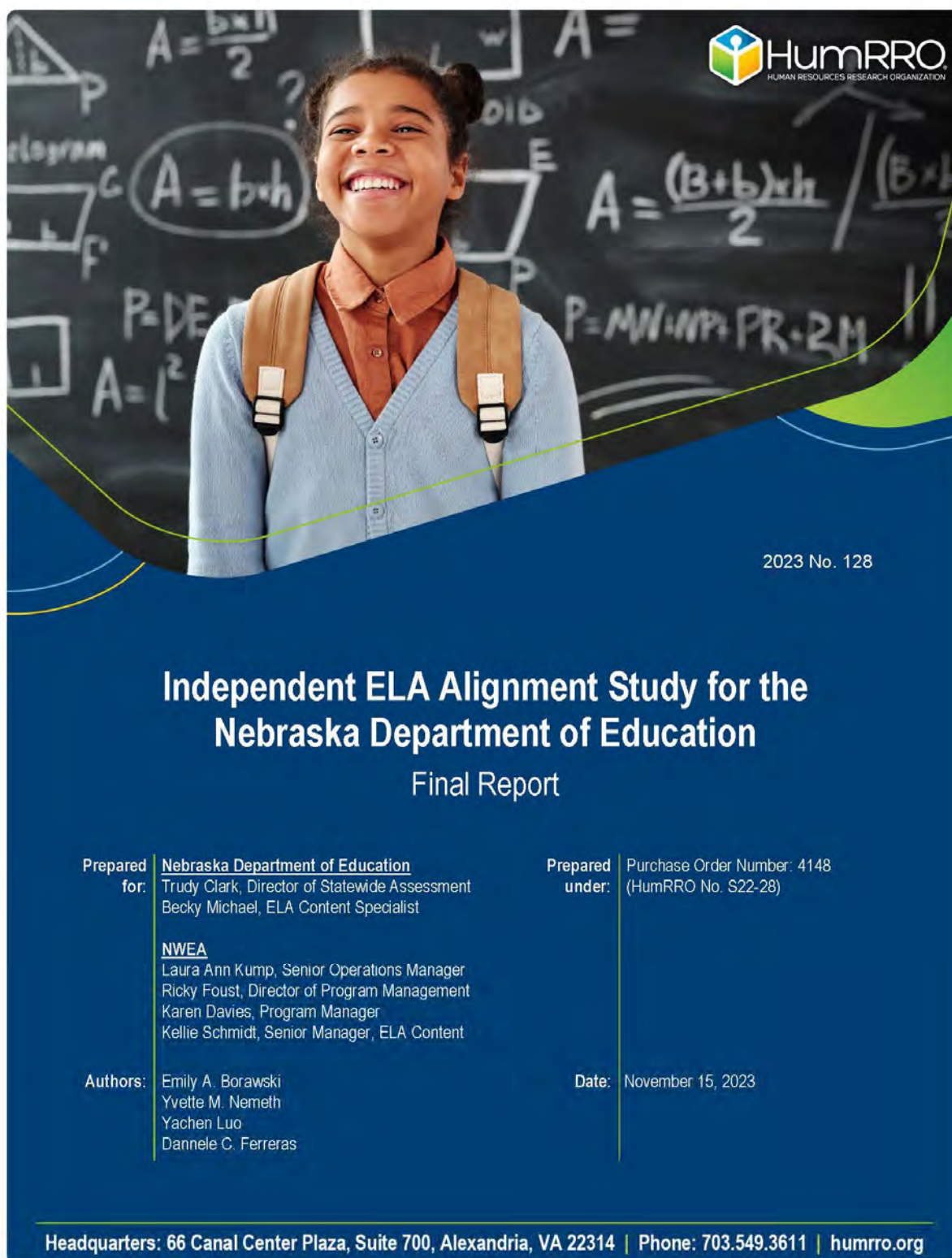


Figure F.3. Scatterplots for Scale Score CSEM—Science

Appendix G: Alignment Study





Independent ELA Alignment Study for the Nebraska Department of Education

Table of Contents

Executive Summary	1
Nebraska Student-Centered Assessment System	1
Nebraska's College and Career Ready Standards	1
Alignment Criteria	1
Test Events	2
Alignment Workshop	2
Overview of Findings	3
Recommendations	5
Criterion 1: Items Represent Intended Content	5
Criterion 2: Items Represent Intended Categories	5
Criterion 3: Depth of Knowledge	5
Criterion 4: Achievement Level Descriptors	6
Chapter 1: Introduction	7
Chapter 2: Methods	9
Nebraska Student-Centered Assessment System	9
Nebraska's College and Career Ready Standards	9
Alignment Criteria	9
Test Events	11
Panelists	11
Facilitator Training	13
Alignment Workshop	13
Test Security	14
Non-Secure Materials	15
Training	15
Chapter 3: Alignment Results	16
Items Assigned to a Nebraska Standard	17
Criterion 1: Items Represent Intended Content	18
Criterion 2: Items Represent Intended Categories	30
Criterion 3: Items Reflect Levels of Cognitive Complexity	44
Criterion 4: Items Reflect Achievement Level Descriptors	51
Process Evaluation Results	58



Table of Contents (Continued)

Chapter 4: Summary and Recommendations	60
Criterion 1: Items Represent Intended Content.....	61
Criterion 2: Items Represent Intended Categories	62
Criterion 3: Depth of Knowledge	63
Criterion 4: Achievement Level Descriptors	63
References.....	64
Appendix A. Agenda	65
Appendix B. Panelist Requirements	67
Appendix C. Panelist Instructions	68
Appendix D. Panelist Training Slides	74
Appendix E. Standards (Grade 3 Example)	79
Appendix F. Cognitive Complexity (DOK Wheel).....	87
Appendix G. Achievement Level Descriptors (Grade 3 Example)	88
Appendix H. Correlation Analysis	89
Appendix I. Number of Unique and Shared Items by Grade, Test Event, and Strand.....	90
Appendix J. DOK by Grade, Test Event, and Strand	94
Appendix K. ALD by Grade, Test Event, and Strand	106
Appendix L. Process Evaluation Tables by Grade.....	118



Table of Contents (Continued)

List of Tables

Table 1. Overall Alignment Benchmark Criteria	3
Table 2. Summary of Results by Criterion and Strand by Grade	4
Table 3. Chapter Descriptions	7
Table 4. Appendix Descriptions	8
Table 5. NSCAS-to-NE Standards Alignment Criteria by Test Event	10
Table 6. Number of Panelists by Grade	11
Table 7. Panelist Demographic Characteristics	12
Table 8. Benchmark Evaluation Criteria by Test Event	16
Table 9. Key Evaluation Documents	17
Table 10. Items Assigned to a Nebraska Standard – All Grades	17
Table 11. Items NOT Assigned to a Nebraska Standard – All Grades	17
Table 12. Criterion 1 Evaluative Benchmark	18
Table 13. Number of Standards Assessed by Test Event and Strand – Grade 3	19
Table 14. Number of Standards Assessed by Test Event and Strand – Grade 4	21
Table 15. Number of Standards Assessed by Test Event and Strand – Grade 5	23
Table 16. Number of Standards Assessed by Test Event and Strand – Grade 6	25
Table 17. Number of Standards Assessed by Test Event and Strand – Grade 7	27
Table 18. Number of Standards Assessed by Test Event and Strand – Grade 8	29
Table 19. Criterion 2 Evaluative Benchmark	30
Table 20. Category Representation Target Percentage Ranges for Test Blueprints and Study Criterion	31
Table 21. Category Representation Percentage Ranges for Study Criterion by Strand and Test Event – Grade 3	33
Table 22. Category Representation Percentage Ranges for Study Criterion by Strand and Test Event – Grade 4	35
Table 23. Category Representation Percentage Ranges for Study Criterion by Strand and Test Event – Grade 5	37
Table 24. Category Representation Percentage Ranges for Study Criterion by Strand and Test Event – Grade 6	39
Table 25. Category Representation Percentage Ranges for Study Criterion by Strand and Test Event – Grade 7	41
Table 26. Category Representation Percentage Ranges for Study Criterion by Strand and Test Event – Grade 8	43
Table 27. Criterion 3 Evaluative Benchmark	44
Table 28. Depth of Knowledge Levels and Definitions	44
Table 29. Distribution of Depth of Knowledge Levels – All Grades	44
Table 30. DOK Assessed by Test Event – Grade 3	45
Table 31. DOK Assessed by Test Event – Grade 4	46
Table 32. DOK Assessed by Test Event – Grade 5	47



Table of Contents (Continued)

Table 33. DOK Assessed by Test Event – Grade 6.....	48
Table 34. DOK Assessed by Test Event – Grade 7.....	49
Table 35. DOK Assessed by Test Event – Grade 8.....	50
Table 36. Achievement Level Descriptor Definitions	51
Table 37. Distribution of Achievement Level Descriptors – All Grades.....	51
Table 38. ALD Assessed by Test Event – Grade 3	52
Table 39. ALD Assessed by Test Event – Grade 4	53
Table 40. ALD Assessed by Test Event – Grade 5	54
Table 41. ALD Assessed by Test Event – Grade 6	55
Table 42. ALD Assessed by Test Event – Grade 7	56
Table 43. ALD Assessed by Test Event – Grade 8	57
Table 45. Overall Alignment Benchmark Criteria	60
Table 46. Summary of Results by Criterion and Strand by Grade Level	61
Table 47. Summary Across Achievement Levels	62
Table H1. Correlation between DOK and ALD by Grade.....	89
Table I1. Number of Unique and Shared Items by Test Event and Strand – Grade 3	91
Table I2. Number of Unique and Shared Items by Test Event and Strand – Grade 4	91
Table I3. Number of Unique and Shared Items by Test Event and Strand – Grade 5	92
Table I4. Number of Unique and Shared Items by Test Event and Strand – Grade 6	92
Table I5. Number of Unique and Shared Items by Test Event and Strand – Grade 7	93
Table I6. Number of Unique and Shared Items by Test Event and Strand – Grade 8	93
Table J1. DOK by Grade, Test Event, and Strand – Grade 3.....	94
Table J2. DOK by Grade, Test Event, and Strand – Grade 4.....	96
Table J3. DOK by Grade, Test Event, and Strand – Grade 5.....	98
Table J4. DOK by Grade, Test Event, and Strand – Grade 6.....	100
Table J5. DOK by Grade, Test Event, and Strand – Grade 7	102
Table J6. DOK by Grade, Test Event, and Strand – Grade 8.....	104
Table K1. ALD by Grade, Test Event, and Strand – Grade 3	106
Table K2. ALD by Grade, Test Event, and Strand – Grade 4	108
Table K3. ALD by Grade, Test Event, and Strand – Grade 5	110
Table K4. ALD by Grade, Test Event, and Strand – Grade 6	112
Table K5. ALD by Grade, Test Event, and Strand – Grade 7	114
Table K6. ALD by Grade, Test Event, and Strand – Grade 8	116
Table L1. Panelist Evaluation Survey Results – Grade 3	118
Table L2. Panelist Evaluation Survey Results – Grade 4	119
Table L3. Panelist Evaluation Survey Results – Grade 5	120
Table L4. Panelist Evaluation Survey Results – Grade 6	121
Table L5. Panelist Evaluation Survey Results – Grade 7	122
Table L6. Panelist Evaluation Survey Results – Grade 8	123



Table of Contents (Continued)

List of Figures

Figure 1. Panelists by County 13

Figure 2. Alignment Workshop Data Collection Steps 14



Independent ELA Alignment Study for the Nebraska Department of Education

Executive Summary

The Northwest Evaluation Association (NWEA), on behalf of the Nebraska Department of Education (NDE), contracted the Human Resources Research Organization (HumRRO) to evaluate the degree of alignment between the Nebraska Student-Centered Assessment System (NSCAS) in English Language Arts (ELA) and Nebraska's College and Career Ready Standards (NE Standards) in ELA. Alignment studies are required as part of the federal assessment peer review process, provide validity evidence that the assessment measures the intended content, and inform future assessment item development. This alignment study gathered critical evidence to support inferences made about students' scores on the NSCAS in ELA.

Nebraska Student-Centered Assessment System

Nebraska's Student-Centered Assessment System (NSCAS) is a "statewide assessment system that embodies Nebraska's holistic view of students. And helps them prepare for success in postsecondary education, career, and civic life" (NSCAS – Nebraska Department of Education, n.d.).

The NSCAS Growth, administered annually in the spring, is the component of NSCAS that assesses whether students have learned what they are expected to learn at their grade level. The test is administered online to all students in Grades 3-8 through Computer Adaptive Testing (CAT). However, a paper-pencil option is available for students with accommodations. The NSCAS in ELA includes approximately 45 test questions and is estimated to take 90 minutes to complete.

Nebraska's College and Career Ready Standards

"Nebraska Revised Statute 79-760.01 requires the Nebraska State Board of Education to 'adopt measurable academic content standards for at least the grade levels required for statewide assessment.' Those standards shall cover the subject areas of reading, writing, mathematics, science, and social studies, and the State Board of Education shall develop a plan to review and update standards for those subject areas every seven years" (Content Area Standards – Nebraska Department of Education, n.d.).

In September 2021, the Nebraska State Board of Education approved Nebraska's College and Career Ready Standards for English Language Arts. The 2021 NE Standards in ELA require students to gain mastery of content in Reading Prose and Poetry (RP), Reading Informational Text (RI), Vocabulary (V), and Writing (W). These content categories will be referred to as "strands" in this report.

Alignment Criteria

Alignment criteria were developed by HumRRO and approved by the Nebraska Department of Education (NDE). The inability to meet all criteria does *not* indicate that the test is invalid, only that a particular assessment aspect may need to be addressed through future item development and modifications to the test specifications.



This alignment study intended to address the following research questions:

1. To what extent does the NSCAS in ELA reflect the breadth of the NE Standards in ELA?
2. To what extent does the NSCAS in ELA reflect the intended distributions of the strands outlined in the test blueprints?
3. To what extent does the NSCAS in ELA reflect a range and distribution of depth of knowledge (DOK)?
4. How well do the Achievement Level Descriptors (ALDs) capture the knowledge and skills expressed in the items?

HumRRO used an alignment methodology based on Webb's original static form alignment criteria (Webb 1997, 1999, 2002, 2005; Wise, et al., 2015). Using this as our base, we tailored the methods to address Nebraska's specific assessment system design for their standards and assessments and current alignment practice. We also applied an aspect of the Achieve model (2018), which incorporates the test blueprints into the alignment evaluation. We collected evidence from the NE Standards, test blueprints, and items from Grades 3-8. The purpose was to gather evidence to support the claims that the assessments align with the test blueprints and that the items are connected to an appropriate NE standard.

Test Events

For the CAT test events, we requested that NWEA randomly select four CAT test events from each of the three achievement levels—Developing, On Track, and Advanced. In each achievement level, the test event was randomly selected from students obtaining the median score within the achievement level score range. Therefore, for each grade level, there were a total of 12 test events (four in Developing, four in On Track, and four in Advanced).

Alignment Workshop

The virtual alignment workshop took place July 24–28, 2023. Based on qualification criteria developed in collaboration with NDE, NWEA, and HumRRO, NWEA recruited educators to serve on grade-level panels in Grades 3-8. Educators participated in a general training session led by HumRRO, which provided background on alignment, an overview of the study's methodology, and item ratings to be collected during the workshop. Panelists received additional training on workshop materials, accessing and navigating the item viewing platform, and data collection processes.

Panelists then performed iterative steps for each item their panels reviewed. These steps included 1.) viewing secure test items, 2.) entering independent ratings into a spreadsheet, 3.) discussing independent ratings with other alignment workshop participants, and 4.) determining final ratings for each item as a group. For final ratings, panelists were instructed to reach a majority agreement (because reaching 100% consensus across all panelists for all items would be too time-consuming for this workshop) on any item in which all panelists disagree with the selected NE Standards, DOK, or ALD. The majority agreement rating for each item on the NE Standard, DOK, and/or ALD was determined through a group discussion by panelists.



Data generated during this study included:

- *Ratings of standard identification, DOK, and ALD.* First, panelists independently identified the NE standard that best captured the assessed item content. Second, panelists were shown the standard to which the item was written. If the independent standard and the intended standard aligned, panelists moved on to their independent DOK rating. If the independent and intended standards did *not* align, panelists identified which standard was a better fit and then moved on to their independent DOK rating. Following the DOK rating, panelists moved on to their independent Achievement Level Descriptor (ALD) rating.
- *Majority ratings of standard identification, DOK, and ALDs.* A majority rating discussion was held for any item that all panelists did not assign the same standard, DOK, or ALD. A customized rating sheet was developed to allow the HumRRO facilitator to record the final majority ratings.
- *Demographic and process evaluation surveys.* At the end of the workshop, panelists completed a process evaluation survey in which they provided feedback about the quality of the workshop. The results of the process evaluation survey are outlined in Appendix L.

Overview of Findings

Table 1 outlines the evaluative guidelines for the overall benchmark criteria, which involves a two-step process. First, test events are evaluated *within* each of the three achievement levels (Developing, On Track, and Advanced). Meeting at least three out of four test event benchmarks results in a "Met" rating while meeting or partially meeting at least two benchmarks leads to a "Partially Met" rating. If fewer than two benchmarks are met or partially met, the criterion is considered "Not Met."

Next, we assess results across the three achievement levels. If all three achievement levels are met, the final criterion is "Met." Meeting or partially meeting two achievement levels leads to a "Partially Met" rating while meeting or partially meeting less than two achievement levels results in a "Not Met" rating. These guidelines offer a structured approach to evaluating and interpreting the overall performance of Criterion 1, 2, and 3 across test events and achievement levels.

Table 1. Overall Alignment Benchmark Criteria

Criteria	Step 1 : Within Achievement Level	Step 2: Across Achievement Levels (Final Rating)
Criterion 1, 2, and 3	<p>Met: At least three out of four test event benchmarks are met within each achievement level.</p> <p>Partially Met: At least two of four test event benchmarks are met or partially met within each achievement level.</p> <p>Not Met: Less than two of four test event benchmarks are met or partially met within each achievement level.</p>	<p>Met: All three achievement levels are met.</p> <p>Partially Met: Two achievement levels are met or partially met.</p> <p>Not Met: Less than two achievement levels are met or partially met.</p>



Table 2 summarizes the alignment criteria results for Grades 3-8.

Criterion 1 measures whether items represent the intended content. Specifically, this criterion measures that alignment between the NE Standards and test items on each test event. For Criterion 1, Reading Prose and Poetry in Grades 3, 5, and 6 partially met the benchmark, while Grades 4, 7, and 8 did not meet the benchmark. Specifically, eight of 12 test events in Grade 4, 8 of 12 test events in Grade 7, and 11 of 12 test events in Grade 8 had less than half of the standards measured by items. For Reading Informational Text and Vocabulary, all grades partially met the benchmark. However, for the Writing strand, the evaluative benchmark was not met across grades. Specifically, 12 of 12 grade test events had less than half of the standards measured by items.

Criterion 2 measures whether items represent intended categories. Specifically, this criterion compares the expected distribution of items by content strand, as presented in the test blueprints, to the distribution of items on each test event. In Criterion 2, all grades partially met or met the benchmark. For Reading Prose and Poetry, Grades 3, 5, 7, and 8 partially met the benchmark, while Grades 4 and 6 met the benchmark. For Reading Informational Text, Grades 3, 4, and 6 partially met, and Grades 5, 7, and 8 met the benchmark. Vocabulary was partially met for Grades 3, 5, and 6 and met for Grades 4, 7, and 8. Lastly, Writing was met for Grades 3, 4, 5, 6, and 8 and partially met for Grade 7.

Criterion 3 measures whether items reflect levels of cognitive complexity. Specifically, the purpose of this criterion is to evaluate the type of cognitive processing required by items to examine the items' breadth of cognitive complexity using Webb's DOK. In Criterion 3, all grades met the benchmark except for Grade 4, which partially met the benchmark.

In summary, while there were variations in performance across different criteria and grade levels, most grades met or partially met the evaluative benchmarks. The results in the body of the report further detail the benchmark criteria by each test event.

Table 2. Summary of Results by Criterion and Strand by Grade

Grade	Criterion 1: Items Represent Intended Content	Criterion 2: Items Represent Intended Categories	Criterion 3: Items Reflect Levels of Cognitive Complexity
Grade 3	RP: Partially Met RI: Partially Met V: Partially Met W: Not Met	RP: Partially Met RI: Partially Met V: Partially Met W: Met	Met
Grade 4	RP: Not Met RI: Partially Met V: Partially Met W: Not Met	RP: Met RI: Partially Met V: Met W: Met	Partially Met
Grade 5	RP: Partially Met RI: Partially Met V: Partially Met W: Not Met	RP: Partially Met RI: Met V: Partially Met W: Met	Met
Grade 6	RP: Partially Met RI: Partially Met V: Partially Met W: Not Met	RP: Met RI: Partially Met V: Partially Met W: Met	Met



Grade	Criterion 1: Items Represent Intended Content	Criterion 2: Items Represent Intended Categories	Criterion 3: Items Reflect Levels of Cognitive Complexity
Grade 7	RP: Not Met RI: Partially Met V: Partially Met W: Not Met	RP: Partially Met RI: Met V: Met W: Partially Met	Met
Grade 8	RP: Not Met RI: Partially Met V: Partially Met W: Not Met	RP: Partially Met RI: Met V: Met W: Met	Met

Recommendations

Criterion 1: Items Represent Intended Content

Based on the results, there is partial support that items represent the intended content. Examination of the blueprint NE Standards to be assessed by items indicates that there are more standards than items allowed, especially with the Writing strand. Based on these findings, we present the following recommendation for NDE's consideration:

- Revise the test specifications to align with the standard level for the Vocabulary and Writing strands rather than the sub-standard level. This is particularly relevant because the Writing strand included 20 or more sub-standards in numerous cases across various grade levels.

Criterion 2: Items Represent Intended Categories

Most benchmarks across grades and content strands were either "Met" or "Partially Met." To strengthen the content strand blueprint target, for any strand that was "Partially Met," we present the following recommendation for NDE's consideration:

- Conduct a review of the NE Standards assigned to items in ELA to ensure Reading Prose and Poetry, Reading Informational Text, and Vocabulary are appropriately associated with the test items. NDE or NWEA can complete this review. Outcomes of this review may include but are not limited to re-assigning an NE Standard to an item.
- Review, across grade-level assessments, the ELA item banks for coverage of content strands. Where necessary, develop more items to ensure an adequate pool for CAT assessments.
- Examine the CAT algorithm to help ensure that the items represent the intended categories specified in the test blueprint.

Criterion 3: Depth of Knowledge

The findings indicate that most items aligned with the DOK level 2. Across all grades, 70% or more of the items were aligned with a DOK level 2 or higher, except for Grade 4. However, there were a handful of grade-levels where no DOK 3 items were administered on one or more test events, while the other test events had at least one DOK 3 item. Based on these findings, we present the following recommendation for NDE's consideration:



- Evaluate the number of DOK 3 items available to determine whether a greater development effort should be made to increase the number of DOK 3 items.
- Continue to ensure balanced and effective item development by focusing on item writing efforts that maintain an appropriate distribution of DOK levels across grade levels.

Criterion 4: Achievement Level Descriptors

The findings indicate that most items aligned with ALD level 2. Across all grades, 70% or more of the items were aligned with an ALD level 2 or higher. However, there were several grade levels where no items were aligned with an ALD level 3. Based on these findings, we present the following recommendation for NDE's consideration:

- Evaluate the number of ALD level 3 items to determine whether a greater development effort should be made to increase the number of ALD level 3 items.
- Continue to ensure balanced and effective item development by focusing on item writing efforts that maintain an appropriate distribution of ALD levels across grade levels.



Chapter 1: Introduction

The Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014) identify alignment as a key component of validity evidence that should be collected for an assessment. Similarly, the federal Assessment Peer Review Guidance specifies that assessments must be aligned to a state's academic content standards (U.S. Department of Education, 2018). Independent alignment studies demonstrate the validity of the assessments based on content. These evaluations document the breadth of knowledge and the level of cognitive processing expected of students during test performance. Alignment results can inform ongoing item development and test form assembly by identifying gaps in content coverage or areas in which the complexity of the test items does not match what is expected of students during instruction. In other words, an alignment study can provide validity evidence about a state assessment system by demonstrating that an assessment (a) represents the full range of the content standards intended to be assessed and (b) measures student knowledge in the same manner and at the same level of complexity as expected in the content standards.

To meet state and Federal requirements, the Nebraska Department of Education (NDE) requested an independent review of the alignment between the Nebraska Standards in English Language Arts (ELA) and Nebraska's Student-Centered Assessment System (NSCAS) in ELA.¹ The Human Resources Research Organization (HumRRO) conducted the requested alignment study in July 2023.

The remaining chapters of this report present detailed information about the methods we used to examine the alignment of the NSCAS with the Nebraska Standards and our analysis of the data we collected.

The chapters are presented as follows:

Table 3. Chapter Descriptions

Chapter	Description
Chapter 1	Chapter 1 provides an introduction to the study and explains the importance of alignment in educational assessments, particularly in relation to validity evidence.
Chapter 2	Chapter 2 explains our alignment method , including the activities we completed to evaluate the alignment of the NSCAS assessment with the Nebraska Standards.
Chapter 3	Chapter 3 presents results describing the alignment of the NSCAS ELA assessment items to standards.
Chapter 4	Chapter 4 provides recommendations for the NDE to strengthen the alignment of the NSCAS assessments over time.

¹ NWEA serves as the vendor for the NSCAS.



Additionally, the appendices are presented as follows:

Table 4. Appendix Descriptions

Appendix	Description
Appendix A	Appendix A contains the Nebraska Alignment Workshop Agenda
Appendix B	Appendix B contains the panelists recruitment requirements
Appendix C	Appendix C contains the panelist rating instructions
Appendix D	Appendix D contains the panelist training slides
Appendix E	Appendix E contains an example of the Grade 3 NE Standards
Appendix F	Appendix F contains the Depth of Knowledge (DOK) Wheel
Appendix G	Appendix G contains an example of the Grade 3 Achievement Level Descriptors (ALD)
Appendix H	Appendix H contains a correlation analysis between DOK and ALD by grade
Appendix I	Appendix I contains the number of unique and shared items by grade, test event, and strand
Appendix J	Appendix J contains DOK ratings by grade, test event, and strand
Appendix K	Appendix K contains ALD ratings by grade, test event, and strand
Appendix L	Appendix L contains the process evaluation results by grade



Chapter 2: Methods

This chapter presents an overview of Nebraska's Student-Centered Assessment System (NSCAS) and Nebraska's College and Career Ready Standards (NE Standards). We also explain our alignment methodology, including the activities we completed to evaluate the alignment of the NSCAS assessment with the NE Standards in ELA.

Nebraska Student-Centered Assessment System

Nebraska's Student-Centered Assessment System (NSCAS) is a "statewide assessment system that embodies Nebraska's holistic view of students. And helps them prepare for success in postsecondary education, career, and civic life" (NSCAS – Nebraska Department of Education, n.d.).

The NSCAS Growth, administered annually in the spring, is the component of NSCAS that assesses whether students have learned what they are expected to learn at their grade level. The test is administered online to all students in Grades 3-8 through Computer Adaptive Testing (CAT); however, a paper-pencil option is available for students with accommodations. The NSCAS in ELA includes approximately 45 test questions and is estimated to take 90 minutes to complete.

Nebraska's College and Career Ready Standards

"Nebraska Revised Statute 79-760.01 requires the Nebraska State Board of Education to 'adopt measurable academic content standards for at least the grade levels required for statewide assessment.' Those standards shall cover the subject areas of reading, writing, mathematics, science, and social studies, and the State Board of Education shall develop a plan to review and update standards for those subject areas every seven years" (*Content Area Standards – Nebraska Department of Education, n.d.*).

In September 2021, the Nebraska State Board of Education approved Nebraska's College and Career Ready Standards for English Language Arts. The 2021 NE Standards in ELA require students to gain mastery of content in Reading Prose and Poetry (RP), Reading Informational Text (RI), Vocabulary (V), and Writing (W). These content categories will be referred to as "strands" in this report.

Alignment Criteria

Alignment studies provide evidence to support the claim that assessments measure the content they are intended to measure. In this case, the content, or the measurement construct, is described for the NSCAS by the 2021 NE Standards in ELA. The alignment workshop was designed to evaluate how well the test items represent (align with) the 2021 NE Standards in ELA. The results presented in this report provide initial evidence of whether the NSCAS ELA assessment measures the content of the NE Standards.

This alignment study intended to address the following research questions:

1. To what extent does the NSCAS in ELA reflect the breadth of the NE Standards in ELA?
2. To what extent does the NSCAS in ELA reflect the intended distributions of the strands outlined in the test blueprints?



3. To what extent does the NSCAS in ELA reflect a range and distribution of depth of knowledge?
4. How well do the Achievement Level Descriptors (ALDs) capture the knowledge and skills expressed in the items?

Our methodology used four criteria to evaluate the alignment of the NSCAS in ELA with the NE Standards in ELA. Table 5 provides a brief description and evaluative benchmark associated with the criteria for each test event.

Table 5. NSCAS-to-NE Standards Alignment Criteria by Test Event

Criteria	Description	Benchmark
Criterion 1: Items Represent Intended Content	This criterion measured the alignment between the NE standards and test items on each test event.	<p>Met: At least 75% of the NE Standards are assessed by items.</p> <p>Partially Met: 50% - 74% of the NE Standards are assessed by items.</p> <p>Not Met: Less than 50% of the NE Standards are assessed by items.</p>
Criterion 2: Items Represent Intended Categories	This criterion compared the expected distribution of items by content strand, as presented in the test blueprints, to the distribution of items on each test event.	<p>Met: Nebraska content strands are $\pm 5\%$ from the minimum and maximum target values outlined in the blueprint.</p> <p>Not Met: Nebraska content strands are not within $\pm 5\%$ of the minimum and maximum target values outlined in the blueprint.</p>
Criterion 3: Items Reflect Levels of Cognitive Complexity	This criterion focused on the cognitive complexity of items. The purpose of this criterion is to evaluate the type of cognitive processing required by items to examine the items' breadth of cognitive complexity using Webb's DOK.	<p>Met: At least 70% of items are rated at cognitive complexity level 2 or above.</p> <p>Not Met: Less than 70% of items are rated at cognitive complexity level 2 or above.</p>
Criterion 4: Items Reflect Levels of Achievement Level Descriptors	This criterion focused on the range of achievement level descriptors (ALDs). Some states include this as additional complexity information in their Peer Review submission. Using well-defined ALDs is consistent with the principles of assessment design.	Items on each test event will reflect a range of ALDs.



Test Events

For the CAT test events, we requested that NWEA, the testing vendor, randomly select four CAT test events from each of the three achievement levels—Developing, On Track, and Advanced. In each achievement level, the test event was randomly selected from students obtaining the median score within the achievement level score range. Therefore, for each grade level, there were a total of 12 test events (four in Developing, four in On Track, and four in Advanced).

Panelists

HumRRO, NWEA, and NDE developed qualification criteria for educators who applied to participate in grade-level review panels for the alignment study. The qualification criteria are presented in Appendix B. Participation requirements were focused on teaching experience and knowledge of the 2021 NE Standards in ELA. NWEA used these qualification criteria to recruit panelists, which were approved by NDE. The number of panelists varied by grade and ranged from three to seven panelists (Table 6).

Table 6. Number of Panelists by Grade

Grade	Number of Panelists
ELA 3	4
ELA 4	3
ELA 5	5
ELA 6	6
ELA 7	5
ELA 8	7

Panelists represented various demographic subgroups and regions across the state of Nebraska. Across all panels, women comprised 100% of the panelists. Most panelists identified as White/Non-Hispanic (90%). Panelists also represented a range of ages, with most panelists between the ages of 26 years old to 55 years old (84%). Additionally, 83% of panelists earned an advanced degree, with 70% having earned a master's degree and 13% having earned a doctoral degree or equivalent. Moreover, panelists were experienced educators, with 57% reporting more than fifteen years of classroom teaching experience. Panelists were also experienced in teaching students from various diverse backgrounds, including but not limited to students from low socioeconomic households (97%), students with disabilities (93%), and English language learners (90%). Table 7 summarizes the demographics of panelists participating in this study.

**Table 7. Panelist Demographic Characteristics**

Category	Description	Count	Percentage
Gender	Woman	30	100%
	Man	0	0%
	Non-Binary	0	0%
Race/Ethnicity	White, Non-Hispanic/Latino	27	90%
	Hispanic/Latino	2	7%
	Prefer not to disclose	1	3%
Age	25 or under	1	3%
	26-35	6	20%
	36-45	8	27%
	46-55	11	37%
	56-65	4	13%
Education	Associate degree	1	3%
	Baccalaureate Degree	4	13%
	Master's Degree	21	70%
	Ph.D. or equivalent (e.g., EdD, JD)	4	13%
Years of Teaching Experience	Under 10	7	23%
	10-14	6	20%
	15 or more	17	57%
Teaching experience with diverse backgrounds*	Yes—Students from low socioeconomic households	29	97%
	Yes—Students receiving free and/or reduced lunch	29	97%
	Yes—Students with disabilities	28	93%
	Yes—English language learners	27	90%
	Yes—Students of color	26	87%
	Other (e.g., students with medical dietary concerns and refugees)	2	3%
	No	1	3%

* Teaching experience with diverse backgrounds is a "select all that apply" response option.
Percentages will sum to greater than 100%.

Panelists represented a variety of counties across the state of Nebraska, with the majority of panelists representing the Douglass (23%) and Lancaster (17%) counties. Figure 1 below provides a visual representation of the counties across the state that were represented by panelists who participated in this study.



Figure 1. Panelists by County



Facilitator Training

In preparation for the alignment workshop, HumRRO led a virtual facilitator training on July 18, 2023, with NDE staff in attendance. The facilitator training focused on providing HumRRO facilitators with an overview of the study background and purpose, workshop materials, alignment basics, data collection process, and facilitator responsibilities.

Alignment Workshop

The virtual alignment workshop took place July 24–28, 2023. Educators participated in a general training session led by HumRRO, which provided background on alignment, an overview of the study's methodology, and the item ratings to be collected during the workshop. Panelists received additional training on workshop materials, accessing and navigating the item viewing platform, and data collection processes.

Panelists then performed iterative steps for each item they reviewed. These steps included 1.) viewing secure test items, 2.) entering independent ratings into a spreadsheet, 3.) discussing independent ratings with other alignment workshop participants, and 4.) determining final ratings for each item as a group. For final ratings, panelists were instructed to reach a majority agreement (because reaching 100% consensus across all panelists for all items would be too time-consuming for this workshop) on any item in which all panelists disagree with the selected Nebraska Standards, DOK, or ALD. The majority agreement rating for each item on the Nebraska Standard, DOK, and/or ALD was determined through a group discussion with all panelists. An overview of these steps is outlined in the graphic below.



Figure 2. Alignment Workshop Data Collection Steps



Data generated during this study included:

- *Ratings of standard identification, DOK, and ALD.* First, panelists independently identified the NE standard that best captured the item content being assessed. Second, panelists were shown the standard to which the item was written. If the independent and intended standards align, panelists moved on to their independent DOK rating. If the independent and intended standards did *not* align, panelists identified which standard was a better fit and then moved on to their independent DOK rating. Following the DOK rating, panelists moved on to their independent Achievement Level Descriptor (ALD) rating.
- *Majority ratings of standard identification, DOK, and ALDs.* A majority rating discussion was held for any item that all panelists did not assign the same standard, DOK, or ALD. A customized rating sheet was developed to allow the HumRRO facilitator to record the final majority ratings.
- *Demographic and process evaluation surveys.* At the end of the workshop, panelists completed demographic and process evaluation surveys in which they provided feedback about the quality of the workshop.

The Grades 4, 5, and 6 panels completed their item ratings early, adjourning on Thursday, July 27, 2023. The Grades 3, 7, and 8 panels completed their item ratings on time and adjourned on Friday, July 28, 2023.

Test Security

Test security was ensured in several ways. First and foremost, all panelists had to sign a non-disclosure agreement (NDA) stating they understood they were responsible for test security of the items being reviewed and would not share any test content with outside individuals. Before the workshop, HumRRO staff were given secure access to the NSCAS ELA items through the on-line Content Review Tool. Accounts for HumRRO facilitators and panelists to log into the Content Review Tool each day during the workshop were created. To further maintain the security of the items, panelist access to the items was turned on each morning and turned off at the conclusion of each workshop day.



Non-Secure Materials

NWEA and NDE also provided HumRRO with several reference materials to help inform panelists' item ratings. These materials included the 2021 NE Standards in ELA (separated by grade level), the Depth of Knowledge Wheel, and the Achievement Level Descriptors (separated by grade level).

In addition to the references provided by NWEA and NDE, HumRRO developed electronic spreadsheets (i.e., Google Sheets) that panelists used to enter item ratings. Facilitators monitored each panelist's ratings in a main spreadsheet. HumRRO also provided training materials, including the panelist and facilitator instructions and training slides for panelists and facilitators. Additionally, HumRRO developed demographic and process evaluation surveys that were used to collect feedback from panelists on demographics, alignment training, and panel facilitation. HumRRO provided all workshop materials to panelists in electronic form through Google Drive.

Training

Panelists participating in the alignment workshop received training before they began rating items. All panelists participated in a general training session led by HumRRO, which provided background on alignment, an overview of the study's alignment methodology and the item ratings to be collected during the workshop.

After the general training session, panelists were released to their grade-level panels for additional in-depth training conducted by their HumRRO facilitator. This training focused on the rating process and the procedures for accessing and using the reference materials to inform their ratings for each item. Panelists then calibrated their ratings with at least the first three items to ensure they shared a common understanding of each rating and used the same approach when evaluating items in the context of the ratings.

NWEA and NDE staff did not engage with panelists beyond the general training session to ensure independence of ratings. However, NWEA and NDE were available to answer panelist questions related to the 2021 NE Standards that HumRRO facilitators communicated via a Microsoft Teams chat and/or video call.



Chapter 3: Alignment Results

This chapter summarizes the data and information collected during the Nebraska ELA alignment workshop. The majority agreement rating for each item was determined through a group discussion. Results are presented for each grade level panel on the following criteria:

Table 8. Benchmark Evaluation Criteria by Test Event

Criteria	Description	Benchmark
Criterion 1: Items Represent Intended Content	This criterion measured the alignment between the NE standards and test items on each test event.	<p>Met: At least 75% of the NE Standards are assessed by items.</p> <p>Partially Met: 50% - 74% of the NE Standards are assessed by items.</p> <p>Not Met: Less than 50% of the NE Standards are assessed by items.</p>
Criterion 2: Items Represent Intended Categories	This criterion compared the expected distribution of items by content strand, as presented in the test blueprints, to the distribution of items on each test event.	<p>Met: Nebraska content strands are $\pm 5\%$ from the minimum and maximum target values outlined in the blueprint.</p> <p>Not Met: Nebraska content strands are not within $\pm 5\%$ of the minimum and maximum target values outlined in the blueprint.</p>
Criterion 3: Items Reflect Levels of Cognitive Complexity	This criterion focused on the cognitive complexity of items. The purpose of this criterion is to evaluate the type of cognitive processing required by items to examine the items' breadth of cognitive complexity using Webb's DOK.	<p>Met: At least 70% of items are rated at cognitive complexity level 2 or above.</p> <p>Not Met: Less than 70% of items are rated at cognitive complexity level 2 or above.</p>
Criterion 4: Items Reflect Levels of Achievement Level Descriptors	This criterion focused on the range of achievement level descriptors (ALDs). Some states include this as additional complexity information in their Peer Review submission. Using well-defined ALDs is consistent with the principles of assessment design.	Items on each test event will reflect a range of ALDs. No specific evaluation benchmark was utilized.

We used four key documents to evaluate the alignment of the NSCAS in ELA with the respective NE Standards:

**Table 9. Key Evaluation Documents**

Key Documents	Description
Nebraska Standards in ELA*	This document lists all standards per grade level in ELA.
Test Blueprints*	This document lists the Nebraska Standards by content strand and the target item percentage by strand.
Cognitive Complexity Definitions	This document provides the cognitive complexity definitions, as defined by Webb.
Achievement Level Descriptors	This document provides the achievement level descriptors (ALDs), which describe the knowledge, skills, and processes that students demonstrate on state tests at pre-determined levels of achievement for each tested grade level.

* It is important to note that the Reading Prose and Poetry (RP) and Reading Informational Text (RI) strands drill down to the standard level and Vocabulary (V) and Writing (W) drill down to the sub-standard level in both the NE Standards and Test Blueprints.

Items Assigned to a Nebraska Standard

Tables 10-11 below describe the number and percentage of items assigned to an NE Standard. The data is disaggregated by grade level, representing how the alignment between items and standards varies across grades. It's worth highlighting that nearly all items assessed content found in the NE Standards across all grade levels, with percentages ranging from 97% to 99%.

Table 10. Items Assigned to a Nebraska Standard – All Grades

Grade	Number of Unique Items	Items Assigned to an NE Standard	
		#	%
ELA 3	246	243	99%
ELA 4	241	234	97%
ELA 5	230	228	99%
ELA 6	215	213	99%
ELA 7	226	223	99%
ELA 8	235	228	97%

Table 11. Items NOT Assigned to a Nebraska Standard – All Grades

Grade	Number of Unique Items	Items NOT Assigned to an NE Standard	
		#	%
ELA 3	246	3	1%
ELA 4	241	7	3%
ELA 5	230	2	1%
ELA 6	215	2	1%
ELA 7	226	3	1%
ELA 8	235	7	3%



Criterion 1: Items Represent Intended Content

Criterion 1 examined the content alignment between items and NE Standards. We reviewed the extent to which items on each of the 12 CAT test events covered the intended NE Standards. For this criterion, we present results evaluating the breadth of NE Standards by grade, test event, and content strand.

Table 12. Criterion 1 Evaluative Benchmark

Benchmark	Description
Met	At least 75% of the NE Standards are assessed by items for each test event.
Partially Met	50%-74% of the NE Standards are assessed by items for each test event.
Not Met	Less than 50% of the NE Standards are assessed by items for each test event.

The NE Standards and test blueprint were the key documents used to evaluate this criterion. First and foremost, the test blueprint lists the content strands of the associated NE Standard that items should measure. The NE Standards are designed and written as discrete statements of the knowledge and skills a student should be taught in each subject and grade level. Locally assessed standards, Foundations of Reading standards, and Speaking and Listening standards were *not* included in the denominators. Additionally, it is important to note that the test blueprint details the Reading Prose and Poetry and Reading Informational Text strands at the *standard level* and the Vocabulary and Writing strands at the *sub-standard level*.

For each grade and each test event, we evaluated the alignment between the items and NE Standards by comparing the number of majority agreement final NE Standards to the number of NE Standards based on the content strands in the test blueprint. Some items were assigned more than one NE Standard by reviewers as the final majority agreement. All assigned NE Standards were included in the counts. As these analyses were based on the majority agreement, items for which reviewers could not identify a NE Standard or a majority agreement could not be reached were excluded from all counts. A detailed breakdown of these data by grade, test event, and blueprint content strands is provided in Tables 13-18. In general, we expected to find that the number of majority agreement NE Standards identified covered the overall range or breadth of blueprint NE Standards listed for each content strand.

In Grade 3, all test events contained items measuring at least half or more of the standards for Reading Prose and Poetry, except for one test event in the Developing achievement level where only three of seven standards were assessed by items. This was also the case for Reading Informational Text except for one test event in the Advanced achievement level where only two of seven standards were assessed by items. For the Vocabulary strand, two test events in the Developing achievement level and two test events in the Advanced achievement level had less than half of the standards measured by items. For the Writing strand, all test events had less than half of the standards measured by items. Across all test events, at least one content strand fell into the Not Met category.



Table 13. Number of Standards Assessed by Test Event and Strand – Grade 3

Achievement Level	Test Event	RP (7 Standards)	RI (7 Standards)	V (6 Standards)	W (20 Standards)	Summary Across Strands
Developing	1	5 (71%) - Partially Met	5 (71%) - Partially Met	2 (33%) - Not Met	4 (20%) - Not Met	2 of 4 – Not Met
	2	3 (43%) - Not Met	5 (71%) - Partially Met	2 (33%) - Not Met	4 (20%) - Not Met	3 of 4 – Not Met
	3	4 (57%) - Partially Met	6 (86%) - Met	4 (67%) - Partially Met	3 (15%) - Not Met	1 of 4 – Not Met
	4	5 (71%) - Partially Met	6 (86%) - Met	3 (50%) - Partially Met	3 (15%) - Not Met	1 of 4 – Not Met
On Track	1	5 (71%) - Partially Met	5 (71%) - Partially Met	2 (33%) - Partially Met	4 (20%) - Not Met	1 of 4 – Not Met
	2	5 (71%) - Partially Met	4 (57%) - Partially Met	3 (50%) - Partially Met	4 (20%) - Not Met	1 of 4 – Not Met
	3	5 (71%) - Partially Met	4 (57%) - Partially Met	3 (50%) - Partially Met	5 (25%) - Not Met	1 of 4 – Not Met
	4	5 (71%) - Partially Met	4 (57%) - Partially Met	3 (50%) - Partially Met	4 (20%) - Not Met	1 of 4 – Not Met
Advanced	1	4 (57%) - Partially Met	5 (71%) - Partially Met	4 (67%) - Partially Met	4 (20%) - Not Met	1 of 4 – Not Met
	2	6 (86%) - Met	5 (71%) - Partially Met	2 (33%) - Not Met	2 (10%) - Not Met	2 of 4 – Not Met
	3	4 (57%) - Partially Met	5 (71%) - Partially Met	2 (33%) - Not Met	4 (20%) - Not Met	2 of 4 – Not Met
	4	4 (57%) - Partially Met	2 (29%) - Not Met	3 (50%) - Partially Met	4 (20%) - Not Met	2 of 4 – Not Met
Summary Across Achievement Levels		1 of 12 – Not Met	1 of 12 – Not Met	4 of 12 – Not Met	12 of 12 – Not Met	–

In Grade 4, three test events in the Developing achievement level, two in the On Track achievement level, and three in the Advanced achievement level had less than half of the standards measured by items for Reading Prose and Poetry. For Reading Informational Text, two test events in the Developing achievement level and one test event in the On Track achievement level had less than half of the standards measured by items. For the Vocabulary strand, one test event from the Developing achievement level, two from the On Track achievement level, and two from the Advanced achievement level had less than half of the standards measured by items. For the Writing strand, all test events had less than half of the standards measured by items. Across all test events, at least one content strand fell into the Not Met category.



Table 14. Number of Standards Assessed by Test Event and Strand – Grade 4

Achievement Level	Test Event	RP (7 Standards)	RI (7 Standards)	V (5 Standards)	W (20 Standards)	Summary Across Strands
Developing	1	3 (43%) - Not Met	4 (57%) - Partially Met	3 (60%) - Partially Met	5 (25%) - Not Met	2 of 4 – Not Met
	2	3 (43%) - Not Met	3 (43%) - Not Met	4 (80%) - Met	6 (30%) - Not Met	3 of 4 – Not Met
	3	5 (71%) - Partially Met	5 (71%) - Partially Met	4 (80%) - Met	3 (15%) - Not Met	1 of 4 – Not Met
	4	2 (29%) - Not Met	2 (29%) - Not Met	2 (40%) - Not Met	5 (25%) - Not Met	4 of 4 – Not Met
On Track	1	5 (71%) - Partially Met	5 (71%) - Partially Met	2 (40%) - Not Met	4 (20%) - Not Met	2 of 4 – Not Met
	2	4 (57%) - Partially Met	5 (71%) - Partially Met	4 (80%) - Met	6 (30%) - Not Met	1 of 4 – Not Met
	3	2 (29%) - Not Met	3 (43%) - Not Met	3 (60%) - Partially Met	6 (30%) - Not Met	3 of 4 – Not Met
	4	3 (43%) - Not Met	4 (57%) - Partially Met	2 (40%) - Not Met	5 (25%) - Not Met	3 of 4 – Not Met
Advanced	1	3 (43%) - Not Met	5 (71%) - Partially Met	3 (60%) - Partially Met	6 (30%) - Not Met	2 of 4 – Not Met
	2	4 (57%) - Partially Met	6 (86%) - Met	2 (40%) - Not Met	4 (20%) - Not Met	2 of 4 – Not Met
	3	3 (43%) - Not Met	4 (57%) - Partially Met	2 (40%) - Not Met	5 (25%) - Not Met	3 of 4 – Not Met
	4	3 (43%) - Not Met	4 (57%) - Partially Met	4 (80%) - Met	5 (25%) - Not Met	2 of 4 – Not Met
Summary Across Achievement Levels		8 of 12 – Not Met	3 of 12 – Not Met	5 of 12 – Not Met	12 of 12 – Not Met	–

In Grade 5, one test event in the v achievement level and two in the Advanced achievement level had less than half of the standards measured by items for Reading Prose and Poetry. For the Reading Informational Text strand, only one test event in the Developing achievement level and one in the Advanced achievement level had less than half of the standards measured by items. For the Vocabulary strand, only one test event from the On Track achievement level had less than half of the standards measured by items. For the Writing strand, all test events had less than half of the standards measured by items. Across all test events, at least one content strand fell into the Not Met category.



Table 15. Number of Standards Assessed by Test Event and Strand – Grade 5

Achievement Level	Test Event	RP (7 Standards)	RI (7 Standards)	V (5 Standards)	W (19 Standards)	Summary Across Strands
Developing	1	5 (71%) - Partially Met	5 (71%) - Partially Met	3 (60%) - Partially Met	3 (16%) - Not Met	1 of 4 – Not Met
	2	5 (71%) - Partially Met	2 (29%) - Not Met	3 (60%) - Partially Met	4 (21%) - Not Met	2 of 4 – Not Met
	3	5 (71%) - Partially Met	6 (86%) - Met	3 (60%) - Partially Met	3 (16%) - Not Met	1 of 4 – Not Met
	4	6 (86%) - Met	5 (71%) - Partially Met	5 (100%) - Met	4 (21%) - Not Met	1 of 4 – Not Met
On Track	1	4 (57%) - Partially Met	4 (57%) - Partially Met	3 (60%) - Partially Met	3 (16%) - Not Met	1 of 4 – Not Met
	2	4 (57%) - Partially Met	5 (71%) - Partially Met	3 (60%) - Partially Met	4 (21%) - Not Met	1 of 4 – Not Met
	3	4 (57%) - Partially Met	6 (86%) - Met	5 (100%) - Met	3 (16%) - Not Met	1 of 4 – Not Met
	4	3 (43%) - Not Met	4 (57%) - Partially Met	2 (40%) - Not Met	3 (16%) - Not Met	3 of 4 – Not Met
Advanced	1	6 (86%) - Met	5 (71%) - Partially Met	4 (80%) - Met	4 (21%) - Not Met	1 of 4 – Not Met
	2	5 (71%) - Partially Met	6 (86%) - Met	4 (80%) - Met	5 (26%) - Not Met	1 of 4 – Not Met
	3	3 (43%) - Not Met	3 (43%) - Not Met	3 (60%) - Partially Met	5 (26%) - Not Met	3 of 4 – Not Met
	4	1 (14%) - Not Met	4 (57%) - Partially Met	4 (80%) - Met	5 (26%) - Not Met	2 of 4 – Not Met
Summary Across Achievement Levels		3 of 12 – Not Met	2 of 12 – Not Met	1 of 12 – Not Met	12 of 12 – Not Met	–



In Grade 6, one test event in the Developing achievement level and one in the On Track achievement level had less than half of the standards measured by items for Reading Prose and Poetry. For the Reading Informational Text strand, only one test event in the Developing achievement level and one in the On Track achievement level had less than half of the standards measured by items. For the Vocabulary strand, only one test event from the Developing achievement level had less than half of the standards measured by items. For the Writing strand, all test events had less than half of the standards measured by items. Across all test events, at least one content strand fell into the Not Met category.



Table 16. Number of Standards Assessed by Test Event and Strand – Grade 6

Achievement Level	Test Event	RP (7 Standards)	RI (7 Standards)	V (5 Standards)	W (22 Standards)	Summary Across Strands
Developing	1	4 (57%) - Partially Met	2 (29%) - Not Met	2 (40%) - Not Met	5 (23%) - Not Met	3 of 4 – Not Met
	2	4 (57%) - Partially Met	5 (71%) - Partially Met	3 (60%) - Partially Met	5 (23%) - Not Met	1 of 4 – Not Met
	3	3 (43%) - Not Met	4 (57%) - Partially Met	4 (80%) - Met	3 (14%) - Not Met	2 of 4 – Not Met
	4	5 (71%) - Partially Met	5 (71%) - Partially Met	4 (80%) - Met	5 (23%) - Not Met	1 of 4 – Not Met
On Track	1	3 (43%) - Not Met	4 (57%) - Partially Met	4 (80%) - Met	5 (23%) - Not Met	2 of 4 – Not Met
	2	5 (71%) - Partially Met	5 (71%) - Partially Met	4 (80%) - Met	4 (18%) - Not Met	1 of 4 – Not Met
	3	6 (86%) - Met	3 (43%) - Not Met	4 (80%) - Met	6 (27%) - Not Met	2 of 4 – Not Met
	4	4 (57%) - Partially Met	5 (71%) - Partially Met	4 (80%) - Met	4 (18%) - Not Met	1 of 4 – Not Met
Advanced	1	4 (57%) - Partially Met	4 (57%) - Partially Met	3 (60%) - Partially Met	4 (18%) - Not Met	1 of 4 – Not Met
	2	5 (71%) - Partially Met	5 (71%) - Partially Met	2 (40%) - Partially Met	4 (18%) - Not Met	1 of 4 – Not Met
	3	6 (86%) - Met	4 (57%) - Partially Met	3 (60%) - Partially Met	4 (18%) - Not Met	1 of 4 – Not Met
	4	4 (57%) - Partially Met	5 (71%) - Partially Met	3 (60%) - Partially Met	3 (14%) - Not Met	1 of 4 – Not Met
Summary Across Achievement Levels		2 of 12 – Not Met	2 of 12 – Not Met	1 of 12 – Not Met	12 of 12 – Not Met	–



In Grade 7, three test events in the Developing achievement level, two in the On Track achievement level, and three in the Advanced achievement level had less than half of the standards measured by items for Reading Prose and Poetry. For the Reading Informational Text strand, only one test event in the On Track achievement level, and one in the Advanced achievement level had less than half of the standards measured by items. For the Vocabulary strand, two test events from the Developing achievement level and all four test events from the On Track achievement level had less than half of the standards measured by items. For the Writing strand, all test events had less than half of the standards measured by items. Across all test events, at least one content strand fell into the Not Met category.



Table 17. Number of Standards Assessed by Test Event and Strand – Grade 7

Achievement Level	Test Event	RP (7 Standards)	RI (7 Standards)	V (5 Standards)	W (20 Standards)	Summary Across Strands
Developing	1	3 (43%) - Not Met	4 (57%) - Partially Met	2 (40%) - Not Met	5 (25%) - Not Met	3 of 4 – Not Met
	2	4 (57%) - Partially Met	4 (57%) - Partially Met	3 (60%) - Partially Met	5 (25%) - Not Met	1 of 4 – Not Met
	3	3 (43%) - Not Met	4 (57%) - Partially Met	3 (60%) - Partially Met	4 (20%) - Not Met	2 of 4 – Not Met
	4	3 (43%) - Not Met	4 (57%) - Partially Met	2 (40%) - Not Met	4 (20%) - Not Met	3 of 4 – Not Met
On Track	1	2 (29%) - Not Met	4 (57%) - Partially Met	2 (40%) - Not Met	4 (20%) - Not Met	3 of 4 – Not Met
	2	4 (57%) - Partially Met	4 (57%) - Partially Met	2 (40%) - Not Met	5 (25%) - Not Met	2 of 4 – Not Met
	3	3 (43%) - Not Met	3 (43%) - Not Met	2 (40%) - Not Met	5 (25%) - Not Met	4 of 4 – Not Met
	4	4 (57%) - Partially Met	4 (57%) - Partially Met	2 (40%) - Not Met	5 (25%) - Not Met	2 of 4 – Not Met
Advanced	1	4 (57%) - Partially Met	5 (71%) - Partially Met	4 (80%) - Met	4 (20%) - Not Met	1 of 4 – Not Met
	2	2 (29%) - Not Met	3 (43%) - Not Met	3 (60%) - Partially Met	4 (20%) - Not Met	3 of 4 – Not Met
	3	2 (29%) - Not Met	4 (57%) - Partially Met	3 (60%) - Partially Met	5 (25%) - Not Met	2 of 4 – Not Met
	4	3 (43%) - Not Met	6 (86%) - Met	4 (80%) - Met	4 (20%) - Not Met	2 of 4 – Not Met
Summary Across Achievement Levels		8 of 12 – Not Met	2 of 12 – Not Met	6 of 12 – Not Met	12 of 12 – Not Met	–



In Grade 8, all test events had less than half of the standards measured by item except for one test event in the On Track achievement level. For the Reading Informational Text strand, two test events in the Developing achievement level, three in the On Track achievement level, and two in the Advanced achievement level had less than half of the standards measured by items. For the Vocabulary strand, one test event from the Developing achievement level and two from the Advanced achievement level had less than half of the standards measured by items. For the Writing strand, all test events had less than half of the standards measured by items. Across all test events, at least two content strands fell into the Not Met category, primarily in the Reading Prose and Poetry and Writing strands.



Table 18. Number of Standards Assessed by Test Event and Strand – Grade 8

Achievement Level	Test Event	RP (7 Standards)	RI (7 Standards)	V (5 Standards)	W (23 Standards)	Summary Across Strands
Developing	1	2 (29%) - Not Met	3 (43%) - Not Met	3 (60%) - Partially Met	5 (22%) - Not Met	3 of 4 – Not Met
	2	1 (14%) - Not Met	4 (57%) - Partially Met	2 (40%) - Not Met	4 (17%) - Not Met	3 of 4 – Not Met
	3	1 (14%) - Not Met	3 (43%) - Not Met	3 (60%) - Partially Met	4 (17%) - Not Met	3 of 4 – Not Met
	4	2 (29%) - Not Met	5 (71%) - Partially Met	3 (60%) - Partially Met	6 (26%) - Not Met	2 of 4 – Not Met
On Track	1	4 (57%) - Partially Met	3 (43%) - Not Met	4 (80%) - Met	5 (22%) - Not Met	2 of 4 – Not Met
	2	3 (43%) - Not Met	3 (43%) - Not Met	3 (60%) - Partially Met	5 (22%) - Not Met	3 of 4 – Not Met
	3	3 (43%) - Not Met	3 (43%) - Not Met	3 (60%) - Partially Met	6 (26%) - Not Met	3 of 4 – Not Met
	4	2 (29%) - Not Met	5 (71%) - Partially Met	3 (60%) - Partially Met	5 (22%) - Not Met	2 of 4 – Not Met
Advanced	1	2 (29%) - Not Met	3 (43%) - Not Met	3 (60%) - Partially Met	6 (26%) - Not Met	3 of 4 – Not Met
	2	3 (43%) - Not Met	4 (57%) - Partially Met	3 (60%) - Partially Met	6 (26%) - Not Met	2 of 4 – Not Met
	3	1 (14%) - Not Met	5 (71%) - Partially Met	2 (40%) - Not Met	4 (17%) - Not Met	3 of 4 – Not Met
	4	2 (29%) - Not Met	1 (14%) - Not Met	1 (20%) - Not Met	6 (26%) - Not Met	4 of 4 – Not Met
Summary Across Achievement Levels		11 of 12 – Not Met	7 of 12 – Not Met	3 of 12 – Not Met	12 of 12 – Not Met	–



Criterion 2: Items Represent Intended Categories

Criterion 2 examined how panelists' majority agreement ratings of items were distributed across content strands. Specifically, we compared the distribution of items using the majority agreement NE Standard compared with the test blueprint target. We generally expected that the majority agreement NE Standard selected for an item would match the content strand targets in the test blueprint.

Table 19. Criterion 2 Evaluative Benchmark

Criteria	Benchmark
Category Representation	<p>Met: Nebraska content strands are +/- 5% from the minimum and maximum target values outlined in the blueprint for each test event.</p> <p>Not Met: Nebraska content strands are not within +/- 5% of the minimum and maximum target values outlined in the blueprint for each test event.</p>

Table 20 presents the target percentage ranges for each strand, based on the test blueprint, and the target percentage ranges for each strand required for this criterion to be met. The target percentage ranges for this study are +/- 5% of the target percentage ranges noted in the test blueprints for each grade.



Table 20. Category Representation Target Percentage Ranges for Test Blueprints and Study Criterion

Grade	RP		RI		V		W	
	Blueprint	Study (+/- 5%)	Blueprint	Study (+/- 5%)	Blueprint	Study (+/- 5%)	Blueprint	Study (+/- 5%)
ELA 3	28%-33%	23%-38%	28%-33%	23%-38%	15%-20%	10%-25%	23%-28%	18%-33%
ELA 4								
ELA 5								
ELA 6								
ELA 7	25%-30%	20%-35%	30%-35%	25%-40%				
ELA 8								



In Grade 3, the Reading Prose and Poetry content strand had two test events in the Developing achievement level that did not meet the blueprint target. The Reading Informational Text content strand had one test event in the Developing achievement level, one in the On Track achievement level, and two in the Advanced achievement level that did not meet the blueprint target. The Vocabulary strand had two test events in the Developing achievement level and one test event in the On Track achievement level that did not meet the blueprint target. Across all three achievement levels, it is noteworthy that all test events successfully met the blueprint target for the Writing content strand.



Table 21. Category Representation Percentage Ranges for Study Criterion by Strand and Test Event – Grade 3

Achievement Level	Test Event	Number of Items	RP (23% - 38%)	RI (23% - 38%)	V (10% - 25%)	W (18% - 33%)	Summary Across Strands
Developing	1	30	27% - Met	40% - Not Met	17% - Met	20% - Met	1 of 4 – Not Met
	2	32	22% - Not Met	31% - Met	28% - Not Met	19% - Met	2 of 4 – Not Met
	3	32	22% - Not Met	34% - Met	25% - Met	19% - Met	1 of 4 – Not Met
	4	31	26% - Met	29% - Met	26% - Not Met	19% - Met	1 of 4 – Not Met
On Track	1	28	32% - Met	21% - Not Met	21% - Met	21% - Met	1 of 4 – Not Met
	2	28	32% - Met	25% - Met	18% - Met	21% - Met	0 of 4 – Not Met
	3	31	26% - Met	32% - Met	26% - Not Met	19% - Met	1 of 4 – Not Met
	4	27	26% - Met	30% - Met	19% - Met	22% - Met	0 of 4 – Not Met
Advanced	1	31	29% - Met	39% - Not Met	13% - Met	19% - Met	1 of 4 – Not Met
	2	28	32% - Met	29% - Met	18% - Met	21% - Met	0 of 4 – Not Met
	3	29	28% - Met	24% - Met	24% - Met	21% - Met	0 of 4 – Not Met
	4	28	36% - Met	18% - Not Met	21% - Met	21% - Met	1 of 4 – Not Met
Summary Across Achievement Levels			2 of 12 – Not Met	4 of 12 – Not Met	3 of 12 – Not Met	0 of 12 – Not Met	–

Note:

- For the On Track achievement level, test events #1, #2, and #4 had one item rated as "None."
- For the Advanced achievement level, test events #3 and #4 had one item rated as "None."



In Grade 4, across all three achievement levels, it is noteworthy that all test events successfully met the blueprint target for the Reading Prose and Poetry, Vocabulary, and Writing content strands. The Reading Informational Text strand had three test events in the Developing achievement level and one test event in the On Track achievement level that did not meet the blueprint target.



Table 22. Category Representation Percentage Ranges for Study Criterion by Strand and Test Event – Grade 4

Achievement Level	Test Event	Number of Items	RP (23% - 38%)	RI (23% - 38%)	V (10% - 25%)	W (18% - 33%)	Summary Across Strands
Developing	1	28	29% - Met	29% - Met	21% - Met	21% - Met	0 of 4 – Not Met
	2	24	29% - Met	17% - Not Met	25% - Met	25% - Met	1 of 4 – Not Met
	3	29	34% - Met	21% - Not Met	24% - Met	21% - Met	1 of 4 – Not Met
	4	27	26% - Met	22% - Not Met	19% - Met	22% - Met	1 of 4 – Not Met
On Track	1	23	30% - Met	30% - Met	13% - Met	26% - Met	0 of 4 – Not Met
	2	28	32% - Met	21% - Not Met	21% - Met	21% - Met	1 of 4 – Not Met
	3	27	26% - Met	30% - Met	19% - Met	22% - Met	0 of 4 – Not Met
	4	27	33% - Met	26% - Met	15% - Met	22% - Met	0 of 4 – Not Met
Advanced	1	27	30% - Met	30% - Met	19% - Met	22% - Met	0 of 4 – Not Met
	2	27	26% - Met	37% - Met	15% - Met	22% - Met	0 of 4 – Not Met
	3	27	26% - Met	33% - Met	22% - Met	22% - Met	0 of 4 – Not Met
	4	29	28% - Met	24% - Met	28% - Met	21% - Met	0 of 4 – Not Met
Summary Across Achievement Levels			0 of 12 – Not Met	4 of 12 – Not Met	0 of 12 – Not Met	0 of 12 – Not Met	—

Note:

- For the Developing achievement level, test event #2 had one item rated as "None."
- For the Developing achievement level, test event #4 had three items rated as "None."
- For the On Track achievement level, test events #2, #3, and #4 had one item rated as "None."
- For the Advanced achievement level, test event #2 had one item rated as "None."



In Grade 5, the Reading Prose and Poetry strand had two test events in the Advanced achievement level that did not meet the blueprint target. The Reading Informational Text strand had one test event in the Developing achievement level and one in the Advanced achievement level that did not meet the blueprint target. The Vocabulary strand had two test events in the On Track achievement level and one test event in the Advanced achievement level that did not meet the blueprint target. Across all three achievement levels, it is noteworthy that all test events successfully met the blueprint target for the Writing content strand.



Table 23. Category Representation Percentage Ranges for Study Criterion by Strand and Test Event – Grade 5

Achievement Level	Test Event	Number of Items	RP (23% - 38%)	RI (23% - 38%)	V (10% - 25%)	W (18% - 33%)	Summary Across Strands
Developing	1	32	31% - Met	31% - Met	19% - Met	19% - Met	0 of 4 – Not Met
	2	25	36% - Met	16% - Not Met	24% - Met	24% - Met	1 of 4 – Not Met
	3	29	24% - Met	34% - Met	21% - Met	21% - Met	0 of 4 – Not Met
	4	32	34% - Met	25% - Met	22% - Met	19% - Met	0 of 4 – Not Met
On Track	1	26	27% - Met	31% - Met	19% - Met	23% - Met	0 of 4 – Not Met
	2	29	28% - Met	31% - Met	24% - Met	21% - Met	0 of 4 – Not Met
	3	30	23% - Met	27% - Met	30% - Not Met	20% - Met	1 of 4 – Not Met
	4	22	32% - Met	27% - Met	9% - Not Met	27% - Met	1 of 4 – Not Met
Advanced	1	29	38% - Met	24% - Met	17% - Met	21% - Met	0 of 4 – Not Met
	2	30	23% - Met	33% - Met	23% - Met	20% - Met	0 of 4 – Not Met
	3	23	22% - Not Met	30% - Met	17% - Met	26% - Met	1 of 4 – Not Met
	4	24	4% - Not Met	42% - Not Met	29% - Not Met	25% - Met	3 of 4 – Not Met
Summary Across Achievement Levels			2 of 12 – Not Met	2 of 12 – Not Met	3 of 12 – Not Met	0 of 12 – Not Met	–

Note.

- For the On Track achievement level, test event #4 had one item rated as "None."
- For the Advanced achievement level, test event #3 had one item rated as "None."



In Grade 6, the Reading Prose and Poetry content strand only had one test event in the Developing achievement level that did not meet the blueprint target. The Reading Informational Text content strand had one test event in the Developing achievement level and two in the On Track achievement level that did not meet the blueprint target. The Vocabulary strand had two test events in the On Track achievement level that did not meet the blueprint target. Across all three achievement levels, it is noteworthy that all test events successfully met the blueprint target for the Writing content strand.



Table 24. Category Representation Percentage Ranges for Study Criterion by Strand and Test Event – Grade 6

Achievement Level	Test Event	Number of Items	RP (23% - 38%)	RI (23% - 38%)	V (10% - 25%)	W (18% - 33%)	Summary Across Strands
Developing	1	23	43% - Not Met	13% - Not Met	17% - Met	26% - Met	2 of 4 – Not Met
	2	28	25% - Met	32% - Met	18% - Met	21% - Met	0 of 4 – Not Met
	3	29	24% - Met	31% - Met	24% - Met	21% - Met	0 of 4 – Not Met
	4	29	28% - Met	31% - Met	21% - Met	21% - Met	0 of 4 – Not Met
On Track	1	28	25% - Met	21% - Not Met	29% - Not Met	21% - Met	2 of 4 – Not Met
	2	30	33% - Met	23% - Met	23% - Met	20% - Met	0 of 4 – Not Met
	3	27	30% - Met	22% - Not Met	26% - Not Met	22% - Met	2 of 4 – Not Met
	4	28	29% - Met	29% - Met	21% - Met	21% - Met	0 of 4 – Not Met
Advanced	1	29	24% - Met	31% - Met	24% - Met	21% - Met	0 of 4 – Not Met
	2	29	24% - Met	31% - Met	24% - Met	21% - Met	0 of 4 – Not Met
	3	26	31% - Met	27% - Met	19% - Met	23% - Met	0 of 4 – Not Met
	4	28	29% - Met	32% - Met	18% - Met	21% - Met	0 of 4 – Not Met
Summary Across Achievement Levels			1 of 12 – Not Met	3 of 12 – Not Met	2 of 12 – Not Met	0 of 12 – Not Met	–

Note:

- For the Developing achievement level, test event #2 had one item rated as "None."
- For the On Track achievement level, test event #1 had one item rated as "None."



In Grade 7, the Reading Prose and Poetry content strand only had two test events in the On Track achievement level that did not meet the blueprint target. Across all three achievement levels, it is noteworthy that all test events successfully met the blueprint target for the Reading Informational Text content strand. The Vocabulary content strand only had one test event in the Developing achievement level that did not meet the blueprint target. The Writing strand had two test events in the Developing achievement level and one test event in the Advanced achievement level that did not meet the blueprint target.



Table 25. Category Representation Percentage Ranges for Study Criterion by Strand and Test Event – Grade 7

Achievement Level	Test Event	Number of Items	RP (20% - 35%)	RI (25% - 40%)	V (10% - 25%)	W (18% - 33%)	Summary Across Strands
Developing	1	31	23% - Met	35% - Met	23% - Met	16% - Not Met	1 of 4 – Not Met
	2	29	24% - Met	31% - Met	24% - Met	21% - Met	0 of 4 – Not Met
	3	28	25% - Met	36% - Met	21% - Met	18% - Met	0 of 4 – Not Met
	4	31	26% - Met	26% - Met	29% - Not Met	16% - Not Met	2 of 4 – Not Met
On Track	1	23	17% - Not Met	39% - Met	17% - Met	26% - Met	1 of 4 – Not Met
	2	27	37% - Not Met	26% - Met	19% - Met	22% - Met	1 of 4 – Not Met
	3	28	25% - Met	36% - Met	18% - Met	21% - Met	0 of 4 – Not Met
	4	27	30% - Met	30% - Met	19% - Met	22% - Met	0 of 4 – Not Met
Advanced	1	27	26% - Met	33% - Met	19% - Met	22% - Met	0 of 4 – Not Met
	2	23	22% - Met	30% - Met	22% - Met	17% - Not Met	1 of 4 – Not Met
	3	27	30% - Met	33% - Met	19% - Met	22% - Met	0 of 4 – Not Met
	4	27	26% - Met	33% - Met	19% - Met	19% - Met	0 of 4 – Not Met
Summary Across Achievement Levels			2 of 12 – Not Met	0 of 12 – Not Met	1 of 12 – Not Met	3 of 12 – Not Met	–

Note.

- For the Developing achievement level, test events #1, #3, and #4 had one item rated as "None."
- For the Advanced achievement level, test event #2 had two items rated as "None."
- For the Advanced achievement level, test event #4 had one item rated as "None."



In Grade 8, the Reading Prose and Poetry content strand had three test events in the Developing achievement level and one in the Advanced achievement level that did not meet the blueprint target. The Reading Informational Text content strand had one test event in the Developing achievement level and one in the On Track achievement level that did not meet the blueprint target. The Vocabulary strand had one test event in the Developing achievement level and one test event in the On Track achievement level that did not meet the blueprint target. Across all three achievement levels, it is noteworthy that all test events successfully met the blueprint target for the Writing content strand.



Table 26. Category Representation Percentage Ranges for Study Criterion by Strand and Test Event – Grade 8

Achievement Level	Test Event	Number of Items	RP (20% - 35%)	RI (25% - 40%)	V (10% - 25%)	W (18% - 33%)	Summary Across Strands
Developing	1	29	38% - Not Met	21% - Met	24% - Met	21% - Met	1 of 4 – Not Met
	2	30	37% - Not Met	30% - Met	13% - Met	23% - Met	1 of 4 – Not Met
	3	24	17% - Not Met	21% - Not Met	25% - Met	25% - Met	2 of 4 – Not Met
	4	32	22% - Met	28% - Met	31% - Not Met	19% - Met	1 of 4 – Not Met
On Track	1	28	32% - Met	25% - Met	18% - Met	21% - Met	0 of 4 – Not Met
	2	22	32% - Met	14% - Not Met	23% - Met	27% - Met	1 of 4 – Not Met
	3	29	24% - Met	31% - Met	24% - Met	21% - Met	0 of 4 – Not Met
	4	30	23% - Met	27% - Met	27% - Not Met	20% - Met	1 of 4 – Not Met
Advanced	1	28	29% - Met	25% - Met	21% - Met	21% - Met	0 of 4 – Not Met
	2	27	26% - Met	33% - Met	19% - Met	22% - Met	0 of 4 – Not Met
	3	25	16% - Not Met	36% - Met	24% - Met	24% - Met	1 of 4 – Not Met
	4	27	30% - Met	26% - Met	15% - Met	22% - Met	0 of 4 – Not Met
Summary Across Achievement Levels			4 of 12 – Not Met	2 of 12 – Not Met	2 of 12 – Not Met	0 of 12 – Not Met	–

Note

- For the Developing achievement level, test event #3 had three items rated as "None."
- For the On Track achievement level, test events #1, #2, and #4 had one item rated as "None."
- For the Advanced achievement level, test event #1 had one item rated as "None."
- For the Advanced achievement level, test event #4 had two items rated as "None."



Criterion 3: Items Reflect Levels of Cognitive Complexity

Criterion 3 is evaluated based on the percentage of items rated by panelists as reflecting each of the three cognitive complexity levels. The criterion is considered "Met" if 70% of items are rated at cognitive complexity level 2 or above. Table 27 provides a brief definition of Webb's DOK Levels.

Table 27. Criterion 3 Evaluative Benchmark

Criterion	Benchmark
DOK Representation	<p>Met: 70% of items are rated at a cognitive complexity level 2 or above.</p> <p>Not Met: Less than 70% of items are rated at a cognitive complexity level 2 or above.</p>

Table 28. Depth of Knowledge Levels and Definitions

Webb's DOK Levels	Definition
Level 1: Recall and Reproduction	Requires recall of information, such as a fact, definition, term, simple procedure, or property. Typically, it involves only one step.
Level 2: Skill/Concept	Requires some mental processing beyond recalling or reproducing a response. Typically, it involves more than one step.
Level 3: Strategic Thinking	Requires deep knowledge using reasoning, planning, or using evidence. Typically, has more than one possible answer and requires students to justify their response.
Level 4: Extended Thinking	Requires high cognitive demand and is very complex. Typically includes complex reasoning, experimental design, and planning, and will likely require an extended period of time.

Table 29 summarizes the number of items and their distribution across the Depth of Knowledge (DOK) levels in Grades 3-8. The data shows that items were predominantly aligned with DOK level 2 across all grades, representing 63% to 85% of the items. DOK 1 represented 5% to 26% of the items, and DOK 3 represented 4% to 26%.

Table 29. Distribution of Depth of Knowledge Levels – All Grades

Grade	Number of Items	DOK 1		DOK 2		DOK 3	
		#	%	#	%	#	%
ELA 3	246	17	7%	170	69%	59	24%
ELA 4	241	62	26%	152	63%	27	11%
ELA 5	230	23	10%	196	85%	10	4%
ELA 6	215	20	9%	156	73%	37	17%
ELA 7	226	12	5%	153	68%	59	26%
ELA 8	235	27	11%	190	81%	17	7%



In Grade 3, all test events met the evaluative benchmark of 70% or more of the items rated at cognitive complexity level 2 or above.

Table 30. DOK Assessed by Test Event – Grade 3

Achievement Level	Test Event	Number of Items	DOK 1	DOK 2	DOK 3	Benchmark
Developing	1	30	1 (3%)	25 (83%)	4 (13%)	Met
	2	32	5 (16%)	21 (66%)	6 (19%)	Met
	3	32	1 (3%)	24 (75%)	7 (22%)	Met
	4	31	3 (10%)	17 (55%)	11 (35%)	Met
On Track	1	28	1 (4%)	18 (64%)	9 (32%)	Met
	2	28	0 (0%)	23 (82%)	5 (18%)	Met
	3	31	2 (6%)	24 (77%)	5 (16%)	Met
	4	27	1 (4%)	22 (81%)	4 (15%)	Met
Advanced	1	31	3 (10%)	21 (68%)	7 (23%)	Met
	2	28	1 (4%)	19 (68%)	8 (29%)	Met
	3	29	0 (0%)	23 (79%)	6 (21%)	Met
	4	28	1 (4%)	23 (82%)	4 (14%)	Met

Note. There was one test item in the Developing 1 and On Track 3 test event that was split between RI and RI; it received a DOK of 2.



In Grade 4, all test events for the On Track and Advanced achievement levels met the evaluative benchmark of 70% of items rated at cognitive complexity level 2 or above. However, test event #2 in the On Track achievement level, while meeting the benchmark, had zero DOK 3 items administered compared to one to seven items on all other test events. Only test event #4 met the evaluative benchmark for the Developing achievement level. Additionally, within the Developing achievement level, test events #1 through #3 displayed a considerably higher alignment with DOK level 1 compared to the other test events.

Table 31. DOK Assessed by Test Event – Grade 4

Achievement Level	Test Event	Number of Items	DOK 1	DOK 2	DOK 3	Benchmark
Developing	1	28	12 (43%)	15 (54%)	1 (4%)	Not Met
	2	24	10 (42%)	13 (54%)	1 (4%)	Not Met
	3	29	10 (34%)	16 (55%)	3 (10%)	Not Met
	4	27	7 (26%)	15 (56%)	5 (19%)	Met
On Track	1	23	6 (26%)	14 (61%)	3 (13%)	Met
	2	28	6 (21%)	22 (79%)	0 (0%)	Met
	3	27	5 (19%)	18 (67%)	4 (15%)	Met
	4	27	5 (19%)	20 (74%)	2 (7%)	Met
Advanced	1	27	3 (11%)	22 (81%)	2 (7%)	Met
	2	27	4 (15%)	18 (67%)	5 (19%)	Met
	3	27	3 (11%)	17 (63%)	7 (26%)	Met
	4	29	7 (24%)	19 (66%)	3 (10%)	Met

Note. There was one test item in the Advanced 2 and Advanced 3 test events that was split between RI and RI₁; it received a DOK of 3.



In Grade 5, all test events met the evaluative benchmark of 70% of items rated at cognitive complexity level 2 or above. However, there were two test events in the Developing achievement level where zero DOK 3 items were administered compared to one or two items on all the other test events.

Table 32. DOK Assessed by Test Event – Grade 5

Achievement Level	Test Event	Number of Items	DOK 1	DOK 2	DOK 3	Benchmark
Developing	1	32	3 (9%)	29 (91%)	0 (0%)	Met
	2	25	2 (8%)	21 (84%)	2 (8%)	Met
	3	29	3 (10%)	24 (83%)	2 (7%)	Met
	4	32	5 (16%)	27 (84%)	0 (0%)	Met
On Track	1	26	2 (8%)	23 (88%)	1 (4%)	Met
	2	29	2 (7%)	26 (90%)	1 (3%)	Met
	3	30	3 (10%)	26 (87%)	1 (3%)	Met
	4	22	1 (5%)	19 (86%)	1 (5%)	Met
Advanced	1	29	2 (7%)	26 (90%)	1 (3%)	Met
	2	30	2 (7%)	26 (87%)	2 (7%)	Met
	3	23	1 (4%)	21 (91%)	1 (4%)	Met
	4	24	2 (8%)	21 (88%)	1 (4%)	Met

Note. For the On Track achievement level in test event #4, there was one item that was "None" for DOK, hence 22 test items.

One test item in the On Track 2 test event was split between RP and RP; it received a DOK of 2.



In Grade 6, all test events met the evaluative benchmark of 70% of items rated at cognitive complexity level 2 or above.

Table 33. DOK Assessed by Test Event – Grade 6

Achievement Level	Test Event	Number of Items	DOK 1	DOK 2	DOK 3	Benchmark
Developing	1	23	1 (4%)	18 (78%)	4 (17%)	Met
	2	28	7 (25%)	17 (61%)	4 (14%)	Met
	3	29	6 (21%)	18 (62%)	5 (17%)	Met
	4	29	4 (14%)	22 (76%)	3 (10%)	Met
On Track	1	28	0 (0%)	23 (82%)	5 (18%)	Met
	2	30	5 (17%)	22 (73%)	3 (10%)	Met
	3	27	0 (0%)	23 (85%)	4 (15%)	Met
	4	28	1 (4%)	18 (64%)	9 (32%)	Met
Advanced	1	29	0 (0%)	22 (76%)	7 (24%)	Met
	2	29	0 (0%)	21 (72%)	8 (28%)	Met
	3	26	0 (0%)	20 (77%)	6 (23%)	Met
	4	28	5 (18%)	17 (61%)	6 (21%)	Met



In Grade 7, all test events met the evaluative benchmark of 70% of items rated at cognitive complexity level 2 or above.

Table 34. DOK Assessed by Test Event – Grade 7

Achievement Level	Test Event	Number of Items	DOK 1	DOK 2	DOK 3	Benchmark
Developing	1	31	4 (13%)	21 (68%)	5 (16%)	Met
	2	29	2 (7%)	21 (72%)	6 (21%)	Met
	3	28	1 (4%)	19 (68%)	7 (25%)	Met
	4	31	2 (6%)	21 (68%)	8 (26%)	Met
On Track	1	23	2 (9%)	11 (48%)	9 (39%)	Met
	2	27	0 (0%)	19 (70%)	8 (30%)	Met
	3	28	2 (7%)	22 (79%)	4 (14%)	Met
	4	27	0 (0%)	18 (67%)	9 (33%)	Met
Advanced	1	27	1 (4%)	18 (67%)	8 (30%)	Met
	2	23	0 (0%)	19 (83%)	4 (17%)	Met
	3	27	0 (0%)	17 (63%)	10 (37%)	Met
	4	27	0 (0%)	20 (74%)	7 (26%)	Met

Notes.

- For the Developing achievement level test event #1, one item was "None" for DOK, hence 31 test items.
- For the Developing achievement level test event #3, one item was "None" for DOK, hence 28 test items.
- For the On Track achievement level test event #1, one item was "None" for DOK, hence 23 test items.
- One test item in the Developing 3 test event was split between RI and RI; it did not receive a DOK rating.
- One test item in the On Track 2 test event was split between RP and V; it received a DOK of 2.
- One test item in the Advanced 3 test event was split between RP and RP; it received a DOK of 3.



In Grade 8, all test events met the evaluative benchmark of 70% of items rated at cognitive complexity level 2 or above. However, there was one test event in the Developing achievement level and one in the On Track achievement level where zero DOK 3 items were administered compared to one to four items on all the other test events.

Table 35. DOK Assessed by Test Event – Grade 8

Achievement Level	Test Event	Number of Items	DOK 1	DOK 2	DOK 3	Benchmark
Developing	1	29	2 (7%)	26 (90%)	1 (3%)	Met
	2	30	5 (17%)	25 (83%)	0 (0%)	Met
	3	24	2 (8%)	17 (71%)	4 (17%)	Met
	4	32	6 (19%)	23 (72%)	3 (9%)	Met
On Track	1	28	2 (7%)	26 (93%)	0 (0%)	Met
	2	22	0 (0%)	18 (82%)	4 (18%)	Met
	3	29	3 (10%)	23 (79%)	3 (10%)	Met
	4	30	2 (7%)	25 (83%)	3 (10%)	Met
Advanced	1	28	3 (11%)	22 (79%)	3 (11%)	Met
	2	27	2 (7%)	22 (81%)	3 (11%)	Met
	3	25	1 (4%)	22 (88%)	2 (8%)	Met
	4	27	5 (19%)	20 (74%)	1 (4%)	Met

Notes.

- For the Developing achievement level test event #3, one item was "None" for DOK, hence 24 test items.
- For the Advanced achievement level test event #4, one item was "None" for DOK, hence 27 test items.
- One test item in the Developing 1 test event was split between RP and RP; it received a DOK of 2.
- One test item in the Developing 2 test event was split between W and W; it received a DOK of 2.

DOK data disaggregated by grade, test event, and strand are located in Appendix J.



Criterion 4: Items Reflect Achievement Level Descriptors

Achievement level descriptors (ALDs) describe the knowledge, skills, and processes that students demonstrate on state tests at pre-determined levels of achievement for each tested grade level. The Nebraska State Board of Education defines three achievement levels:

1. Developing
2. On Track
3. Advanced

Table 36. Achievement Level Descriptor Definitions

ALD Levels	Definition
Level 1: Developing	Developing learners do not yet demonstrate proficiency in the knowledge and skills necessary at this grade level, as specified in the assessed Nebraska College and Career Ready Standards. These results prove that the student may need additional support for academic success at the next grade level.
Level 2: On Track	On Track learners demonstrate proficiency in the knowledge and skills necessary at this grade level, as specified in the assessed Nebraska College and Career Ready Standards. These results prove that the student will likely be ready for academic success at the next grade level.
Level 3: Advanced	Advanced learners demonstrate proficiency in the knowledge and skills necessary at this grade level, as specified in the assessed Nebraska College and Career Ready Standards. These results prove that the student will likely be ready for academic success at the next grade level.

Table 37 describes the number of items and their distribution across the ALDs in Grades 3-8. The data shows that across all grades, items were predominantly aligned with ALD level 2, representing 58% to 80% of the items. ALD 1 represented 8% to 27% of the items, and ALD 3 represented 2% to 26%.

Table 37. Distribution of Achievement Level Descriptors – All Grades

Grade	Number of Items	ALD 1		ALD 2		ALD 3	
		#	%	#	%	#	%
ELA 3	246	20	8%	161	65%	64	26%
ELA 4	241	66	27%	139	58%	29	12%
ELA 5	230	42	18%	183	80%	4	2%
ELA 6	215	39	18%	157	73%	16	7%
ELA 7	226	47	21%	138	61%	40	18%
ELA 8	235	39	17%	172	73%	21	9%



In Grade 3, items were classified into the three categories as follows: ALD 1 (ranging from 3% to 19%), ALD 2 (ranging from 46% to 77%), and ALD 3 (ranging from 16% to 43%).

Table 38. ALD Assessed by Test Event – Grade 3

Achievement Level	Test Event	Number of Items	ALD 1	ALD 2	ALD 3
Developing	1	30	2 (7%)	23 (77%)	5 (17%)
	2	32	6 (19%)	21 (66%)	5 (16%)
	3	32	1 (3%)	22 (69%)	9 (28%)
	4	31	2 (6%)	23 (74%)	6 (19%)
On Track	1	28	3 (11%)	13 (46%)	12 (43%)
	2	28	3 (11%)	20 (71%)	5 (18%)
	3	31	3 (10%)	23 (74%)	5 (16%)
	4	27	1 (4%)	20 (74%)	6 (22%)
Advanced	1	31	1 (3%)	20 (65%)	10 (32%)
	2	28	3 (11%)	15 (54%)	10 (36%)
	3	29	1 (3%)	19 (66%)	9 (31%)
	4	28	2 (7%)	20 (71%)	5 (18%)

Note. For the Advanced achievement level test event #4, one item was "None," hence 28 test items.

One test item in the Developing 1 and On Track 3 test events was split between RI and RI; it received an ALD of 2.



In Grade 4, items were classified into the three categories as follows: ALD 1 (ranging from 19% to 39%), ALD 2 (ranging from 52% to 70%), and ALD 3 (ranging from 4% to 19%).

Table 39. ALD Assessed by Test Event – Grade 4

Achievement Level	Test Event	Number of Items	ALD 1	ALD 2	ALD 3
Developing	1	28	9 (32%)	17 (61%)	2 (7%)
	2	24	8 (33%)	14 (58%)	1 (4%)
	3	29	8 (28%)	18 (62%)	3 (10%)
	4	27	7 (26%)	14 (52%)	3 (11%)
On Track	1	23	9 (39%)	12 (52%)	2 (9%)
	2	28	8 (29%)	16 (57%)	3 (11%)
	3	27	7 (26%)	14 (52%)	5 (19%)
	4	27	8 (30%)	14 (52%)	4 (15%)
Advanced	1	27	9 (33%)	16 (59%)	2 (7%)
	2	27	5 (19%)	19 (70%)	2 (7%)
	3	27	5 (19%)	18 (67%)	4 (15%)
	4	29	7 (24%)	18 (62%)	4 (14%)

Notes:

- For the Developing achievement level test event #2, one item was "None," hence 24 test items.
- For the Developing achievement level test event #4, three items were "None," hence 27 test items.
- For the On Track achievement level test event #2, there was one item that was "None," hence 28 test items.
- For the On Track achievement level test event #3, one item was "None," hence 27 test items.
- For the On Track achievement level test event #4, one item was "None," hence 27 test items.
- For the Advanced achievement level test event #2, one item was "None," hence 27 test items.
- One test item in the Advanced 2 and Advanced 3 test events was split between RI and RI; it received an ALD of 2.



In Grade 5, items were classified into the three categories as follows: ALD 1 (ranging from 7% to 28%), ALD 2 (ranging from 72% to 93%), and ALD 3 (ranging from 0% to 5%). Overall, there were few items assigned to ALD 3 across all test events.

Table 40. ALD Assessed by Test Event – Grade 5

Achievement Level	Test Event	Number of Items	ALD 1	ALD 2	ALD 3
Developing	1	32	4 (13%)	28 (88%)	0 (0%)
	2	25	4 (16%)	21 (84%)	0 (0%)
	3	29	7 (24%)	21 (72%)	1 (3%)
	4	32	7 (22%)	25 (78%)	0 (0%)
On Track	1	26	4 (15%)	22 (85%)	0 (0%)
	2	29	7 (24%)	22 (76%)	0 (0%)
	3	30	2 (7%)	28 (93%)	0 (0%)
	4	22	4 (18%)	16 (73%)	1 (5%)
Advanced	1	29	8 (28%)	21 (72%)	0 (0%)
	2	30	4 (13%)	25 (83%)	1 (3%)
	3	23	6 (26%)	17 (74%)	0 (0%)
	4	24	4 (17%)	19 (79%)	1 (4%)

Note. For the On Track achievement level test event #4, there was one item that was "None," hence 22 test items. One test item in the On Track 2 test event was split between RP and RP; it received an ALD of 1.



In Grade 6, items were classified into the three categories as follows: ALD 1 (ranging from 4% to 36%), ALD 2 (ranging from 61% to 92%), and ALD 3 (ranging from 0% to 21%). Two test events were in the Developing achievement level where no items were assigned an ALD 3.

Table 41. ALD Assessed by Test Event – Grade 6

Achievement Level	Test Event	Number of Items	ALD 1	ALD 2	ALD 3
Developing	1	23	3 (13%)	20 (87%)	0 (0%)
	2	28	10 (36%)	17 (61%)	0 (0%)
	3	29	8 (28%)	19 (66%)	2 (7%)
	4	29	6 (21%)	21 (72%)	2 (7%)
On Track	1	28	3 (11%)	22 (79%)	3 (11%)
	2	30	4 (13%)	24 (80%)	2 (7%)
	3	27	2 (7%)	24 (89%)	1 (4%)
	4	28	1 (4%)	21 (75%)	6 (21%)
Advanced	1	29	5 (17%)	23 (79%)	1 (3%)
	2	29	4 (14%)	21 (72%)	4 (14%)
	3	26	1 (4%)	24 (92%)	1 (4%)
	4	28	5 (18%)	21 (75%)	2 (7%)

Note. For the Developing achievement level test event #2, there was one item that was "None," hence 28 test items.



In Grade 7, items were classified into the three categories as follows: ALD 1 (ranging from 11% to 42%), ALD 2 (ranging from 45% to 76%), and ALD 3 (ranging from 10% to 30%).

Table 42. ALD Assessed by Test Event – Grade 7

Achievement Level	Test Event	Number of Items	ALD 1	ALD 2	ALD 3
Developing	1	31	13 (42%)	14 (45%)	4 (13%)
	2	29	4 (14%)	22 (76%)	3 (10%)
	3	28	5 (18%)	15 (54%)	7 (25%)
	4	31	9 (29%)	15 (48%)	7 (23%)
On Track	1	23	4 (17%)	12 (52%)	7 (30%)
	2	27	6 (22%)	17 (63%)	4 (15%)
	3	28	6 (21%)	19 (68%)	3 (11%)
	4	27	6 (22%)	18 (67%)	3 (11%)
Advanced	1	27	5 (19%)	16 (59%)	6 (22%)
	2	23	3 (13%)	14 (61%)	6 (26%)
	3	27	3 (11%)	18 (67%)	6 (22%)
	4	27	4 (15%)	20 (74%)	3 (11%)

Notes.

- For the Developing achievement level test event #3, one item was "None," hence 28 test items.
- One test item in the Developing 3 test event was split between RI and RI; it did not receive an ALD rating.
- One test item in the On Track 2 test event was split between RP and V; it received an ALD of 3.
- One test item in the Advanced 3 test event was split between RP and RP; it received an ALD of 3.



In Grade 8, items were classified into the three categories as follows: ALD 1 (ranging from 4% to 33%), ALD 2 (ranging from 63% to 89%), and ALD 3 (ranging from 0% to 20%). One test event in the On Track achievement level was assigned no ALD 3 items.

Table 43. ALD Assessed by Test Event – Grade 8

Achievement Level	Test Event	Number of Items	ALD 1	ALD 2	ALD 3
Developing	1	29	5 (17%)	22 (76%)	2 (7%)
	2	30	10 (33%)	19 (63%)	1 (3%)
	3	24	4 (17%)	18 (75%)	2 (8%)
	4	32	7 (22%)	23 (72%)	2 (6%)
On Track	1	28	2 (7%)	25 (89%)	0 (0%)
	2	22	2 (9%)	16 (73%)	3 (14%)
	3	29	3 (10%)	24 (83%)	2 (7%)
	4	30	5 (17%)	23 (77%)	2 (7%)
Advanced	1	28	1 (4%)	22 (79%)	4 (14%)
	2	27	1 (4%)	23 (85%)	3 (11%)
	3	25	2 (8%)	18 (72%)	5 (20%)
	4	27	5 (19%)	19 (70%)	3 (11%)

Notes.

- For the On Track achievement level test event #1, one item was "None," hence 28 test items.
- For the On Track achievement level test event #2, one item was "None," hence 22 test items.
- For the Advanced achievement level, test event #1 had one item that was "None," hence 28 test items.
- One test item in the Developing 1 test event was split between RP and RP; it received an ALD of 1.
- One test item in the Developing 2 test event was split between W and W; it received an ALD of 2.

ALD data disaggregated by grade, test event, and strand are located in Appendix K.



Process Evaluation Results

Upon adjourning each panel, HumRRO facilitators administered a process evaluation survey to their panelists.

Overall, panelists evaluated the workshop with high levels of satisfaction (Table 44). On a scale of 1-5, with 1 = "Strongly Disagree" to 5 = "Strongly Agree," most panelists believed their facilitator did an effective job of facilitating discussion and ensuring all panelists' perspectives were heard (average = 4.90), the facilitators clearly and promptly addressed questions (average = 4.90), and the facilitator was helpful during the workshop (average = 4.83). Notably, across all panels, 50% of educators reported strong alignment of items with the NE Standards, while the remaining 50% reported partial alignment.

Appendix L provides the complete results of this survey disaggregated by grade level.



Table 44. Panelist Evaluation Survey Results – All Grades

Question	Average	
My panel facilitator did an effective job of facilitating discussion and ensuring that all panelists' perspectives were heard	4.90	<div></div>
My panel facilitator clearly and promptly addressed my questions	4.90	<div></div>
The panel facilitator was helpful during the workshop	4.83	<div></div>
My ideas and opinions were listened to and respected by the group	4.77	<div></div>
Everyone had equal opportunity to contribute ideas and opinions	4.77	<div></div>
The materials hosted on Google Drive were useful (e.g., standards)	4.73	<div></div>
Practicing making ratings as a group in my assigned panel helped me better understand the alignment activities	4.73	<div></div>
The hands-on training in my assigned panel was an effective use of time	4.70	<div></div>
The panel-specific hands-on training was well organized	4.63	<div></div>
The other materials shared by my facilitator were useful	4.63	<div></div>
The Google Rating Sheet was useful for recording alignment ratings	4.60	<div></div>
The Google Rating Sheet provided a comprehensive platform for capturing alignment of standards	4.60	<div></div>
It was easy to access the evaluation and demographics forms	4.60	<div></div>
The whole-group training facilitator was helpful during the workshop	4.57	<div></div>
The hands-on training in my assigned panel helped me better understand the alignment activities	4.50	<div></div>
Other support staff were helpful during the workshop	4.47	<div></div>
The group-wide training session was well organized	4.41	<div></div>
The group-wide training session provided a useful overview of the alignment activities for the week	4.37	<div></div>
The group-wide training session effectively outlined the purpose of the alignment workshop	4.37	<div></div>
The group-wide training session clearly described my role as a panelist	4.37	<div></div>
The Content Review Tool allowed me to effectively accomplish my tasks during the workshop	4.37	<div></div>
The group-wide training was an effective use of time	4.33	<div></div>
It was easy to access the items on the Content Review Tool	4.13	<div></div>

Note: Strongly Disagree = 1 to Strongly Agree = 5



Chapter 4: Summary and Recommendations

In this chapter, we provide an overview of the benchmark criteria, a summary of findings, and recommendations for NDE to consider based on these results. For ease of organization, the summary and recommendations are presented separately for each alignment criterion.

Table 45 outlines the evaluative guidelines for the overall benchmark criteria, which involves a two-step process. First, test events are evaluated *within* the three achievement levels (Developing, On Track, and Advanced). Meeting at least three out of four test event benchmarks results in a "Met" rating while meeting or partially meeting at least two benchmarks leads to a "Partially Met" rating. If fewer than two benchmarks are "Met" or "Partially Met," the criterion is considered "Not Met."

Next, we assess results *across* the three achievement levels. If all three achievement levels are met, the final criterion is "Met." Meeting or partially meeting two achievement levels leads to a "Partially Met" rating while meeting or partially meeting less than two achievement levels results in a "Not Met" rating. These guidelines offer a structured approach to evaluating and interpreting the overall performance of Criterion 1, 2, and 3 across test events and achievement levels by grade.

Table 45. Overall Alignment Benchmark Criteria

Criteria	Step 1 : Within Achievement Level	Step 2: Across Achievement Levels (Final Rating)
Criterion 1, 2, and 3	<p>Met: At least three out of four test event benchmarks are met within each achievement level.</p> <p>Partially Met: At least two of four test event benchmarks are met or partially met within each achievement level.</p> <p>Not Met: Less than two of four test event benchmarks are met or partially met within each achievement level.</p>	<p>Met: All three achievement levels are met.</p> <p>Partially Met: Two achievement levels are met or partially met.</p> <p>Not Met: Less than two achievement levels are met or partially met.</p>

Table 46 summarizes the alignment criteria results for the NSCAS ELA assessments for Grades 3-8.



Table 46. Summary of Results by Criterion and Strand by Grade Level

Grade	Criterion 1	Criterion 2	Criterion 3
Grade 3	RP: Partially Met RI: Partially Met V: Partially Met W: Not Met	RP: Partially Met RI: Partially Met V: Partially Met W: Met	Met
Grade 4	RP: Not Met RI: Partially Met V: Partially Met W: Not Met	RP: Met RI: Partially Met V: Met W: Met	Partially Met
Grade 5	RP: Partially Met RI: Partially Met V: Partially Met W: Not Met	RP: Partially Met RI: Met V: Partially Met W: Met	Met
Grade 6	RP: Partially Met RI: Partially Met V: Partially Met W: Not Met	RP: Met RI: Partially Met V: Partially Met W: Met	Met
Grade 7	RP: Not Met RI: Partially Met V: Partially Met W: Not Met	RP: Partially Met RI: Met V: Met W: Partially Met	Met
Grade 8	RP: Not Met RI: Partially Met V: Partially Met W: Not Met	RP: Partially Met RI: Met V: Met W: Met	Met

Criterion 1: Items Represent Intended Content

This criterion examined the content alignment between 12 test events and the NE Standards. Specifically, we reviewed the majority agreement of the NE Standard identified for each item on the 12 test events.

The results show a diverse range of alignment between the test items and the standards outlined in the test blueprint. Examining the extent to which test events across grades met the Criterion 1 Benchmark, most test events by grade and strand "Partially Met" this criterion (Table 46). However, consistent across all grade levels is the recurring issue of test items not covering the breadth of writing standards outlined in the test blueprint. Many test items intended to evaluate writing proficiency consistently fell short of covering the number of writing standards.

A noteworthy finding is the dynamic nature of alignment with standards as students advance to higher grade levels. For example, Table 47 shows that Grade 3 did not meet benchmark criteria for one of 12 test events for Reading Prose and Poetry, one of 12 for Reading Informational Text, and four of 12 for Vocabulary. However, Grade 4 did not meet benchmark criteria for eight of 12 test events for Reading Prose and Poetry, three of 12 for Reading Informational Text, and five of 12 for Vocabulary—an increase in "Not Met" ratings across all three content strands. Additionally, Table 47 shows that Grade 6 did not meet benchmark criteria for two of 12 test



events for Reading Prose and Poetry, two of 12 for Reading Informational Text, and one of 12 for Vocabulary. However, Grades 7 and 8 did not meet benchmark criteria for eight of 12 and 11 of 12 test events for Reading Prose and Poetry, two of 12 and 7 of 12 for Reading Informational Text, and six of 12 and three of 12 for Vocabulary (respectively)—an increase in “Not Met” ratings across content strands.

Table 47. Summary Across Achievement Levels

Grade	RP	RI	V	W
Grade 3	1 of 12 – Not Met	1 of 12 – Not Met	4 of 12 – Not Met	12 of 12 – Not Met
Grade 4	8 of 12 – Not Met	3 of 12 – Not Met	5 of 12 – Not Met	12 of 12 – Not Met
Grade 5	3 of 12 – Not Met	2 of 12 – Not Met	1 of 12 – Not Met	12 of 12 – Not Met
Grade 6	2 of 12 – Not Met	2 of 12 – Not Met	1 of 12 – Not Met	12 of 12 – Not Met
Grade 7	8 of 12 – Not Met	2 of 12 – Not Met	6 of 12 – Not Met	12 of 12 – Not Met
Grade 8	11 of 12 – Not Met	7 of 12 – Not Met	3 of 12 – Not Met	12 of 12 – Not Met

Based on the results, there is partial support that items represent the intended content. Examination of the blueprint NE Standards to be assessed by items indicates that there are more standards than items allowed, especially with the Writing strand. Based on these findings, we present the following recommendation for NDE's consideration:

- Revise the test specifications to align with the Standard level for the Vocabulary and Writing Strands rather than the sub-standard level. This is particularly relevant because the Writing strand included 20 or more sub-standards in numerous cases across various grade levels.

Criterion 2: Items Represent Intended Categories

This criterion examined how items on each test event met the test blueprint targets for each content strand. Across grades and content strands, most benchmarks were either “Met” or “Partially Met” (Table 46). To strengthen the content strand blueprint target, we recommend the following for any strand that was “Partially Met.”

- Conduct a review of the NE Standards assigned to items in ELA to ensure Reading Prose and Poetry, Reading Informational Text, and Vocabulary are appropriately associated with the test items. This review can be completed by NDE or NWEA. Outcomes of this review may include but are not limited to re-assigning an NE Standard to an item.
- Review, across grade-level assessments, the ELA item banks for coverage of content strands. Where necessary, develop more items to ensure an adequate pool to draw from for CAT assessments.
- Examine the CAT algorithm to help ensure that the items represent the intended categories specified in the test blueprint.



Criterion 3: Depth of Knowledge

This criterion assessed the depth of knowledge of items. We examined the number of items at each DOK level across items on each test event using majority agreement DOK ratings.

Overall, the findings indicate that most items aligned with the DOK level 2. Across all grades, 70% or more of the items were aligned with a DOK level 2 or higher, except for three Developing test events in Grade 4. However, there were a handful of test events where no DOK 3 items were administered, specifically one test event Grade 4 On Track, two test events Grade 5 Developing, one test event Grade 8 Developing and one test event Grade 8 On Track. All other test events had at least one DOK 3 item. Based on these findings, we present the following recommendation for NDE's consideration:

- Evaluate the number of DOK 3 items available to determine whether a greater development effort should be made to increase the number of DOK 3 items.
- Continue to ensure balanced and effective item development by focusing on item writing efforts that maintain an appropriate distribution of DOK levels across grade levels.

Criterion 4: Achievement Level Descriptors

This criterion assessed the range of achievement level descriptors of items. We examined the number of items at each ALD level on each test event using majority agreement ALD ratings.

Overall, the findings indicate that most items aligned with ALD level 2. Across all grades, 70% or more of the items were aligned with an ALD level 2 or higher. However, there were several grade levels where no items were aligned with an ALD level 3. Based on these findings, we present the following recommendation for NDE's consideration:

- Evaluate the number of ALD level 3 items to determine whether a greater development effort should be made to increase the number of ALD level 3 items.
- Continue to ensure balanced and effective item development by focusing on item writing efforts that maintain an appropriate distribution of ALD levels across grade levels.



References

- Achieve, Inc. (2018). *Criteria for procuring and evaluating high-quality and aligned summative science assessments*. Retrieved from <https://www.nextgenscience.org/sites/default/files/Criteria03202018.pdf>
- American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. American Educational Research Association.
- Content Area Standards* - Nebraska Department of Education. (n.d.). <https://www.education.ne.gov/assessment/nscas-system/>
- NSCAS Overview* - Nebraska Department of Education. (n.d.). <https://www.education.ne.gov/assessment/nscas-system/>
- U.S. Department of Education. (2018). A state's guide to the U.S. Department of Education's Assessment Peer Review Process.
- Webb, N. L. (1997). *Research Monograph No. 6: Criteria for alignment of expectations and assessments in Science and Science education*. Council of Chief State Schools Officers.
- Webb, N. L. (1999). *Research Monograph No. 18: Alignment of Science and Science standards and assessments in four states*. National Institute for Science Education and Council of Chief State School Officers. (ERIC Document Reproduction Service No. ED440852).
- Webb, N. L. (2002). *Alignment study in language arts, mathematics, science, and social studies of state standards and assessments for four states*. Council of Chief State School Officers.
- Webb, N. L. (2005). *Webb alignment tool: Training manual*. Wisconsin Center for Education Research. (n.d.). <http://www.wcer.wisc.edu/WAT/index.aspx>
- Wise, S. L., Kingsbury G. G., & Webb, N. L. (2015). Evaluating content alignment in computerized adaptive testing. *Educational Measurement: Issues and Practice*, 00, 1-8.



Appendix A. Agenda

Nebraska Student-Centered Assessment System (NSCAS) in ELA Virtual Alignment Workshop July 24 – July 28, 2023 Agenda

Note: All times noted on the agenda are Central Time

Date/Time	Description
Day 1 – Monday, July 24, 2023	
8:30 a.m. – 10:00 a.m.	Join Microsoft Teams meeting with all panelists and HumRRO Facilitators. Welcome, logistics, overview of NSCAS in ELA, general alignment training
10:00 a.m. – 10:15 a.m.	Break
10:15 a.m. – 11:45 a.m.	Join Teams meeting for assigned grade level panel, panelist introductions, confirm access to online documents and Content Review Tool for NSCAS in ELA, review panelist instructions for rating items and calibrate item ratings, and begin iterative alignment rating process
11:45 a.m. – 12:45 p.m.	Lunch Break
12:45 p.m. – 2:15 p.m.	Continue iterative alignment rating process
2:15 p.m. – 2:30 p.m.	Break
2:30 p.m. – 4:00 p.m.	Continue iterative alignment rating process
4:00 p.m.	<i>Adjourn for the day</i>
Day 2 – Tuesday, July 25, 2023	
8:30 a.m. – 10:00 a.m.	If needed: Review and rerate items from Day 1. Continue iterative alignment rating process
10:00 a.m. – 10:15 a.m.	Break
10:15 a.m. – 11:45 a.m.	Continue iterative alignment rating process
11:45 a.m. – 12:45 p.m.	Lunch Break
12:45 p.m. – 2:15 p.m.	Continue iterative alignment rating process
2:15 p.m. – 2:30 p.m.	Break
2:30 p.m. – 4:00 p.m.	Continue iterative alignment rating process
4:00 p.m.	<i>Adjourn</i>



Date/Time	Description
Day 3 – Wednesday, July 26, 2023	
8:30 a.m. – 10:00 a.m.	If needed: Review and rerate items from Day 2. Continue iterative alignment rating process
10:00 a.m. – 10:15 a.m.	Break
10:15 a.m. – 11:45 a.m.	Continue iterative alignment rating process
11:45 a.m. – 12:45 p.m.	Lunch Break
12:45 p.m. – 2:15 p.m.	Continue iterative alignment rating process
2:15 p.m. – 2:30 p.m.	Break
2:30 p.m. – 4:00 p.m.	Continue iterative alignment rating process
4:00 p.m.	Adjourn
Day 4 – Thursday, July 27, 2023	
8:30 a.m. – 10:00 a.m.	If needed: Review and rerate items from Day 3. Continue iterative alignment rating process
10:00 a.m. – 10:15 a.m.	Break
10:15 a.m. – 11:45 a.m.	Continue iterative alignment rating process
11:45 a.m. – 12:45 p.m.	Lunch Break
12:45 p.m. – 2:15 p.m.	Continue iterative alignment rating process
2:15 p.m. – 2:30 p.m.	Break
2:30 p.m. – 4:00 p.m.	Continue iterative alignment rating process
4:00 p.m.	Adjourn
Day 5 – Friday, July 28, 2023	
8:30 a.m. – 10:00 a.m.	If needed: Review and rerate items from Day 3. Continue iterative alignment rating process
10:00 a.m. – 10:15 a.m.	Break
10:15 a.m. – 11:45 a.m.	Continue iterative alignment rating process
11:45 a.m. – 12:45 p.m.	Lunch Break
12:45 p.m. – 2:15 p.m.	Continue iterative alignment rating process
2:15 p.m. – 2:30 p.m.	Break
2:30 p.m. – 3:45 p.m.	Continue iterative alignment rating process
3:45 p.m. – 4:00 p.m.	Complete two short online surveys: • Demographic information • Debrief/ Workshop evaluation
4:00 p.m.	Adjourn



Appendix B. Panelist Requirements

Nebraska teachers will serve as panelists for the alignment workshop. All reviewers will be confirmed by NDE.

Educators will have the following minimum qualifications for serving as a reviewer:

- Educators are **seasoned, certified, professionals**, who have **strong familiarity** with the Nebraska Standards in ELA. Educators
- Educators are current teachers with at least three years of teaching experience at their respective grade level or teachers who retired after 2021 when the standards were adopted
- Educators have at least read the Nebraska ELA Standards for their grade and related grade span
- Educators have participated in professional development activities related to the Nebraska Standards in ELA, including prior participation with reviewing test items (e.g., prior alignment study experience, prior standard setting study experience)
- Educators have developed a curriculum that incorporates the Nebraska Standards in ELA
- Educators have experience with the NE standards and Range ALDs



Appendix C. Panelist Instructions

Nebraska ELA Alignment Workshop Panelist Instructions

#	Title of Material
1	Panelist instructions
2	Panelist rating sheets
3	Panelist training slides
4	Nebraska ELA items - Accessed via the Content Review Tool
5	Nebraska's College and Career Ready Standards for English Language Arts (NE Standards)
6	Depth of Knowledge (DOK) Levels (Cognitive Complexity)
7	Achievement Level Descriptors (ALDs)
8	Demographics form (via MS Forms) – administered at the end of the workshop
9	Process evaluation survey (via MS Forms) – administered at the end of the workshop

Terminology:

- **NE Standards:** Nebraska's College and Career Ready Standards for English Language Arts
- **NSCAS:** Nebraska's Student-Centered Assessment System (NSCAS)

Test Security Notice

Please do not use your personal electronic devices while engaged in alignment workshop tasks. If you need to use your phone or other devices for any reason, please step away from the computer or wait to use your devices during a break.
This rule will be strictly enforced during the workshop.



Task 1: Introductions and Materials Overview

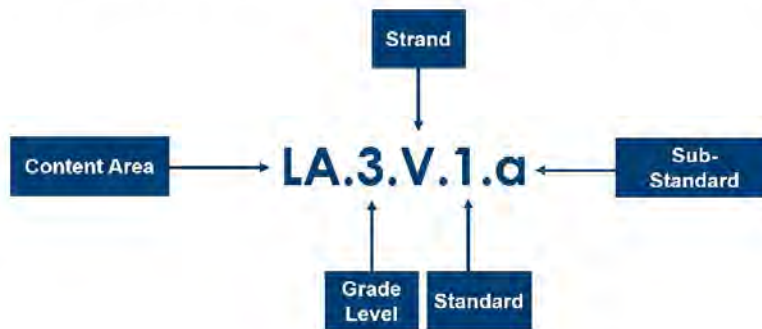
- **Panelist and facilitator introductions**
- **Review the materials in the table above**
 - Google Drive folder with digital materials
 - Facilitator demonstration of how to access Google Sheets. (Follow along on your computer.)
 - Facilitator demonstration of how to log on to the Content Review Tool. (Follow along on your computer.)
 - Materials #8 and #9 (Demographics form and Process Evaluation survey) will be administered at the end of the workshop.

Task 2: Training on the NSCAS ELA item alignment

- **Brief explanation of the process for this task.**
 - You will review NSCAS ELA items administered to students in Nebraska.
 - You will first calibrate your ratings by reviewing a small set of items (typically, the first three to five items). This will be an opportunity for the group to talk through the process and everyone discuss their approach to reviewing each item. This will ensure everyone is thinking about the ratings in the same way. You will then enter your ratings into your rating sheet using the drop-down menus. You will assign a NE Standard(s) that **best** match what the item measures. Your ratings will focus on the alignment of each item to content within the NE Standards, cognitive complexity (DOKs), and achievement level descriptors (ALDs). We will discuss each of these ratings and settle on a final majority rating.
- After calibration, you will independently review a small set of items and enter your ratings into your rating sheet per the instructions above. You will review items in logical sets (e.g., all items in a cluster). Once all items in a set have been reviewed, we will discuss the items as a group and settle on a final majority rating.
- Below is a graphic from the general panelist training that provides a high-level overview of the process.



- Below is a graphic that lays out the structure of the NE Standard codes:



The NE Standards contain the following components:

1. Content area ("LA" refers to Language Arts)
 2. Grade Level
 3. Strand:
 - a. RP = Reading Prose and Poetry (standard level)
 - b. RI = Reading for Informational Text (standard level)
 - c. V = Vocabulary (sub-standard level)
 - d. W = Writing (sub-standard level)
 4. Standard (always a number)
 5. Sub-standard (always a lowercase letter, V & W only)
- Open individual rating sheets (Google Sheets). Open Google Chrome on your computer and navigate to your individual Google rating sheet. Each sheet has a unique panelist name.
 - Review making ratings in the Google Sheet.



Facilitator demonstration on how to enter data in the sheet (i.e., using drop-down menus, entering comments).

- You will need to review only the first sheet. Other sheets are hidden and should not be accessed or modified. If any issue occurs with the drop-down menu options or conditional formatting, notify your facilitator.

Discuss **Columns A and B (Item Sequence and UIN)**

- Columns A and B contain information about each NSCAS ELA item.
- **Column A** indicates the order the item appears in the Content Review Tool. This number will be what you and the panelists use to make sure everyone is talking about the same item.
- **Column B** provides the unique item number (UIN). You will not use this for your ratings, but it is provided in case any items seem to be out of sequence in the Content Review Tool.
- Please ensure that you are viewing the same item in the Content Review Tool that you are rating in your sheet.

Discuss **Column C (Identify the Standard)**

- Column C asks you to identify the Standard code using a drop-down menu.
- You should be very familiar with the 2021 NE Standards document from which the Standard codes are derived. You are permitted to use your own marked-up copies of the 2021 NE Standards if you have their own.
- **If you believe the item does not align with a NE Standard, you should select "None."** You must enter a comment explaining the reason you entered "None" for this rating in **Column K**.
- If you have a difficult time choosing between two or more Standards, you should select the Standard that **best aligns** with the item. Then, you should enter a comment in **Column K** that includes the other Standard(s) you considered.

Discuss **Column D (Item Writer's Standard)**

- Once you select a Standard code from **Column C**, the Standard code associated with the item per the item writer will appear in **Column D**. The purpose of this is for you to see what Standard you selected and compare it with the assigned Standard from the item writer.

Discuss **Columns E and F (Standard Text)**

- **Column E** will display the Standard text associated with the selected Standard Code in **Column C**.
- **Column F** will display the Standard text associated with the metadata Standard in **Column D**.

Discuss **Column G (Final Rating)**

- **Column G** will ask you to choose between your selected Standard in **Column C** and the item writer's Standard in **Column D**.



Discuss Column H (Final Standard Rating Description)

- **Column H** will ask you to briefly describe if your final Standard in **Column G** differs from the item writer's Standard in **Column D**, and to please indicate why.

Discuss Column I (Identify the Depth of Knowledge/ Cognitive Complexity Level)

- **Column I** is for you to provide the overall cognitive complexity level (1, 2, 3, or 4) that best represents the cognitive demand of the item. Remember you'll need to evaluate the cognitive complexity and **not** the item difficulty (although highly correlated, they are not always the same). Keep in mind that cognitive complexity refers to what the item is asking the student to do and how that task fits into the cognitive complexity framework (reference the DOK document provided).

Discuss Column J (Identify the Range Achievement Level Descriptor)

- **Column J** is for you to provide the range achievement level descriptor (1 - Developing, 2 – On Track, or 3 - Advanced). You'll want to reference the Range ALD document provided.

Discuss Column K (Comments)

- **Column K** is for you to enter any comments.
- A few simple rules for the comments field:
 - a. If the comments cell is highlighted yellow, it means one or more of the following ratings were selected: "None" for the final Standard **OR** you selected a Standard that differed from the item writer's Standard.
 - b. You may also provide comments or notes regarding the quality of the item or the phenomenon the item references. Panelists should take notes on their own, discuss them, and the facilitator should capture the agreed-upon points in the facilitator spreadsheet.
 - c. The primary purpose of this column is to provide comments related to the alignment of the item to the rating categories. All comments will be **anonymously** provided to the Nebraska Department of Education for review.

Task 3: Rating Calibration Task

- You will rate all indicated fields for the **first item**. Since this is a calibration activity, you should read the item, review the reference materials, then work together to come up with a rating for each rating category. The calibration is a collaborative activity, though you should be reminded that, after calibration, you will rate items independently, then discuss their ratings with the rest of the panel once the rest of the panel has finished rating a set of items.
- During calibration, you should focus on why you agree or disagree and what the most appropriate selections should be. Be sure you spend a little time with cognitive complexity and achievement level descriptors.
- You will repeat calibration for up to two to four additional items.

**Task 4: Conduct Independent Item Ratings**

- You should rate all remaining NSCAS ELA items independently in sets before discussing and settling on a majority (items are typically rated in sets based on the corresponding passages). Repeat the process above for each set of items. You will review items in clusters so that any linked items are not broken into separate review and rating sessions.
- You will work independently; however, occasional discussion about any item(s) that is causing someone difficulty is allowed.
- **After discussing an item, you should not change your rating unless you made a coding error.** The facilitator will capture majority ratings among the panelists, but HumRRO wants to be able to gauge the differences between independent panelist ratings and the final majority ratings.

Task 5: Workshop Debrief

- Once all final majority ratings have been collected, please close all materials (e.g., rating sheet, Content Review Tool, any electronic versions of references) and open the MS Word document with the link to the debriefing surveys.
- You will first take the "Demographic Survey" followed by the "Process Evaluation Survey."
- Please note that your responses will be anonymous and will only be shared in an aggregate format.



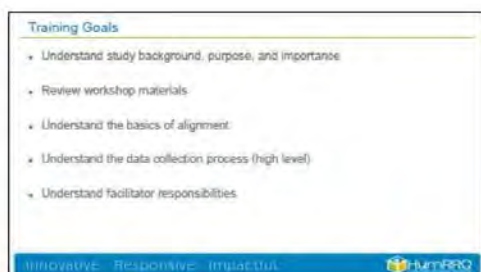
Appendix D. Panelist Training Slides



1



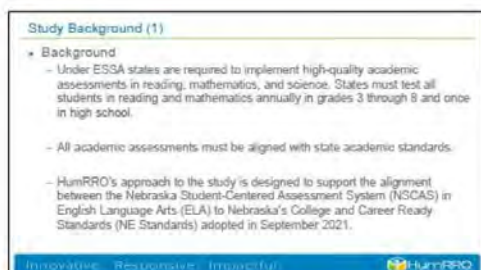
2



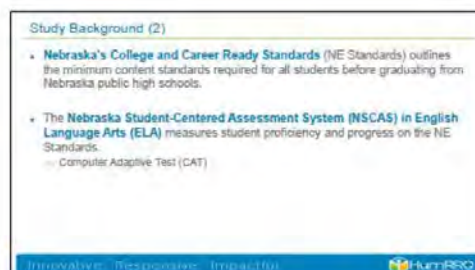
3



4



5



6



Purpose

- Purpose
 - To collect evidence of the alignment of the NSCAS in ELA to the NE Standards (validity evidence)
 - To evaluate the content representation of the test events
 - To evaluate the range and distribution of cognitive complexity or depth of knowledge (DOK)
 - To evaluate how well the achievement level descriptors (ALDs) capture the knowledge and skills expressed in the item

Innovative... Responsive... Impactful

7

Importance of Alignment

- Fairness for all students
 - Consistency in general curriculum
 - Accurate assessment of what students can do and are expected to know from State content standards and the curriculum
 - Improves teacher instruction and student learning
- Federal Peer Review Requirement

Innovative... Responsive... Impactful

8

Workshop Materials

Innovative... Responsive... Impactful

9

Workshop Materials – Google Drive (1)

#	Title of Material
1	Facilitator instructions
2	Facilitator rating sheet
3	Facilitator training slides
4	Panelist instructions
5	Panelist rating sheets
6	Panelist training slides

Innovative... Responsive... Impactful

10

Workshop Materials – Google Drive (2)

#	Title of Material
7	NE ELA Items – Accessed via the Content Review Tool
8	NE Standards
9	Cognitive complexity (DOK levels)
10	Achievement level descriptors (ALDs)
11	Demographics form (MS Form)
12	Process evaluation survey (MS Form)

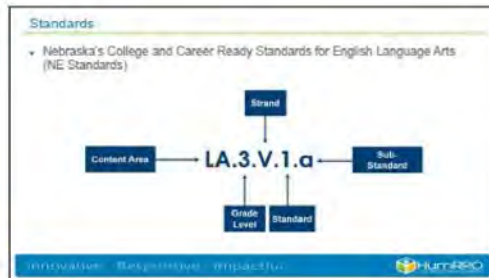
Innovative... Responsive... Impactful

11

**Standards,
Cognitive Complexity
(DOK), and Achievement
Level Descriptors (ALDs)**

Innovative... Responsive... Impactful

12



13

Cognitive Complexity

- Refers to the type of **cognitive processing** required to access and respond to the test item
- Frequently measured using Webb's Depth of Knowledge (DOK) definitions
- Cognitive complexity is related to but distinct from difficulty

Innovative · Responsive · Impactful

14

Cognitive Complexity: Webb's Depth of Knowledge Levels

Webb's DOK Level	Definition
Level 1: Recall and Reproduction	Requires recall of information, such as a fact, definition, term, simple procedure, or property. Typically involves only one step.
Level 2: Skill/Concept	Requires some mental processing beyond recalling or reproducing a response. Typically involves more than one step.
Level 3: Strategic Thinking	Requires deep knowledge using reasoning, planning, or using evidence. Typically has more than one possible answer and requires students to justify their response.
Level 4: Extended Thinking (this term is not aligned with DOK 4)	Requires high cognitive demand and is very complex. Typically involves complex reasoning, experiential design and planning, and probably will require an extended period.

Innovative · Responsive · Impactful

15

Achievement Level Descriptors

- Achievement level descriptors (ALDs) describe the knowledge, skills, and processes that students demonstrate on state tests at pre-determined levels of achievement for each tested grade level.
- The Nebraska State Board of Education defined three achievement levels for each content area:
 - 1. **Developing**
 - 2. **On Track**
 - 3. **Advanced**

Innovative · Responsive · Impactful

16

Achievement Level Descriptors

- 1. Developing:** Developing learners do not yet demonstrate proficiency in the knowledge and skills necessary at this grade level, as specified in the Nebraska College and Career Ready Standards.
- 2. On Track:** On Track learners demonstrate proficiency in the knowledge and skills necessary at this grade level, as specified in the assessed Nebraska College and Career Ready Standards.
- 3. Advanced:** Advanced learners demonstrate proficiency in the knowledge and skills necessary at this grade level, as specified in the assessed Nebraska College and Career Ready Standards.

Innovative · Responsive · Impactful

17

Virtual Alignment Workshop


Innovative · Responsive · Impactful

18



Virtual Alignment Workshop

- We will collect judgments from Nebraska educators on the following:
 - Items to standards alignment
 - Items to cognitive complexity alignment (DOK)
 - Items to achievement level descriptors (ALDs)

INNOVATIVE • RESPONSIVE • IMPACTFUL 

19

The Alignment Rule

- For items, "aligned" means the item's content is included in the standard
- Most items cannot capture the full breadth of a standard
- Think of the Standards as buckets – If the item fits in any of these buckets, it is aligned. (It does not need to fill the bucket.)

INNOVATIVE • RESPONSIVE • IMPACTFUL 

20

Data Collection Process (1)

- Virtual Meeting using multiple platforms/tools
 - Microsoft Teams
 - Google Drive
 - Item Viewing Platform (Content Review Tool)
- Six (6) grade-level panels (Grades 3 through 8)
- Up to 7.5 hours per day, over five (5) days, including scheduled breaks
 - July 24 – July 28, 8:30 am – 4:00 pm Central Time
- Start with whole-group training and then break into panels
 - EB will send out the whole group training invite (8:30am – 10:00am CT)
 - Facilitators will send separate meeting info next (10:15am – 4:00pm CT, 8:30am – 4:00pm CT)

INNOVATIVE • RESPONSIVE • IMPACTFUL 

21

Data Collection Process (2)


- Prior to the workshop
 - EB will send facilitators an email with a link to the Google Drive folder
- Whole-group training
 - EB will send out the whole-group training invite from HumRRO
 - Setup background and purpose
 - Overview of the alignment method
- Small group sessions
 - EB will send facilitators an email with instructions on sending your panelists two MS Teams meeting invites for your individual panels

INNOVATIVE • RESPONSIVE • IMPACTFUL 

22

Data Collection Process (3)

- What is to be viewed/rated
 - Test items
 - Accessed via the Content Review Tool (CRT)
- Materials to support panelist item ratings
 - Panelist instructions
 - Panelist rating sheet
 - Panelist training slides
 - NE Standards
 - Cognitive complexity levels (DOK levels)
 - Achievement level descriptors (ALDs)

INNOVATIVE • RESPONSIVE • IMPACTFUL 

23

Data Collection Process (4)




INNOVATIVE • RESPONSIVE • IMPACTFUL 

24



Facilitation Scenario (2)


- Scenario: The majority of panelists independently select the same rating.
- Is this during calibration (first three items)?
 - Yes: Have panelists discuss ratings.
 - No: Does this involve a standard/cognitive complexity rating/achievement level descriptor that the group has not discussed yet?
 - Yes: Have panelists discuss ratings.
 - No: Record majority rating.

Innovative... Responsive... Impactful. 

31

Facilitation Scenario (3)

- Scenario: Panelists are evenly split on a standard, cognitive complexity, or achievement level descriptor.
- Is this a standard rating?
 - Yes: If the tie cannot be broken, record "split" as the final standard rating and note in the facilitator comments the standards the group was split on.
- Is this a cognitive complexity or achievement level descriptor rating?
 - Yes: If the ratings are non-adjacent, have group discuss further to ensure common understanding of the levels and possibly get to a majority rating. If still evenly split, record the higher DCK or ALD level and comment in the facilitator rating form that this was an evenly split rating and reference the lower rating.

Innovative... Responsive... Impactful. 

32

HumRRO Support Staff

- Emily Borowski:
 - Whole-Group Training Facilitator
 - Floater
- Yvette Nemeth:
 - Technical Advisor
 - Floater
- Camie Wiley:
 - Floater

Innovative... Responsive... Impactful. 

33

Thank You!

Charge Code:
922-08 NE ELA&MathAlignmt

For any questions,
please contact:
Emily Borowski
eborowski@humro.org



Innovative... Responsive... Impactful.  humro.org

34



Appendix E. Standards (Grade 3 Example)

NEBRASKA'S COLLEGE AND CAREER READY
STANDARDS FOR ENGLISH LANGUAGE ARTS



Approved by the Nebraska State Board of Education on September 2, 2021





K-12 Comprehensive English Language Arts Standards

Strand	Comprehensive Standard
Foundations of Reading (F)	Students will develop and apply decoding and language comprehension skills and strategies to comprehend and learn from increasingly complex texts.
Reading Prose and Poetry (RP)	Students will learn and apply reading skills and strategies to comprehend grade-level literary texts.
Reading Informational Text (R)	Students will learn and apply reading skills and strategies to comprehend grade-level informational texts.
Vocabulary (V)	Students will build and use conversational, academic, and discipline-specific, grade-level vocabulary.
Writing (W) and Foundations of Writing (FW)	Students will learn and apply writing skills and strategies to communicate effectively for a variety of purposes.
Speaking and Listening (SL)	Students will learn and apply speaking and listening skills and strategies to communicate effectively for a variety of audiences and purposes.

Spiraled, Vertical Progressions. The revised 2021 Nebraska English Language Arts Standards are formatted to support educators in both grade-level and vertical instructional planning. In addition to organization by grade level, the standards and indicators are formatted into spiraled, vertical articulations. This design demonstrates the interrelated nature of skills in the English Language Arts and their progression through the grade levels. The purpose of presenting the standards into vertical charts is to provide educators with a practical tool for the development of a locally-determined, standards-aligned curriculum.

For each standard in the areas of Foundations of Reading, Reading Prose and Poetry, Reading Informational Text, Writing*, Vocabulary, and Speaking and Listening, the standards and indicators are listed in a table format from the K-12 grade band and ending at Kindergarten.

Page 7

Approved by the Nebraska State Board of Education on September 2, 2021



Grade 3 Standards



READING PROSE AND POETRY

Central Ideas and Details | Citing relevant and thorough textual evidence to support ideas, evaluate the development of themes or central ideas in grade-level literary texts.

LA.3.RP.1 Identify the central message or lesson in a literary text and explain how key details support that idea.

LA.3.RP.2 Explain how characters respond to major events and challenges in a literary text.

Author's Craft | Citing relevant and thorough evidence to support ideas, evaluate the development and interaction of individuals, ideas, and events in grade-level literary and informational text.

LA.3.RP.3 Determine and explain the point of view in a literary text.

LA.3.RP.4 Explain how sections of a literary text (e.g., chapters, scenes, stanzas) build on one another and contribute to meaning.

Knowledge and Ideas | Citing relevant and thorough textual evidence to support ideas, evaluate how an author's perspective or use of point of view shapes the style and meaning of grade-level literary text.

LA.3.RP.5 Compare and contrast the themes, settings, and plots of literary texts written by the same author about the same or similar characters (e.g., books from a series).

LA.3.RP.6 Explain what the text says explicitly and draw inferences when asking and answering questions.

LA.3.RP.7 Compare and contrast themes, topics, and/or patterns of events in a range of literary texts.

Do not select the following standard - locally assessed only:

Range of Reading and Level of Text Complexity | Read and comprehend complex, grade-level literary text independently and proficiently.

LA.3.RP.8 Read and comprehend a wide range of literary texts of appropriate complexity for Grade 3 independently and proficiently.

Instructional Considerations

- In describing settings or characters, students should explain what in the text the descriptions are based upon.
- Students should be made aware that most narratives contain a central message or lesson.
- At all grade levels, students should read paired, conceptually-related (by topic, theme, and/or genre) literary and informational texts.
- Point of view refers to the vantage point from which a narrative is told.

Page 47

Approved by the Nebraska State Board of Education on September 2, 2021



READING INFORMATIONAL TEXT

Central Ideas and Details | Citing relevant and thorough textual evidence to support ideas, evaluate the development of themes or central ideas in grade-level informational text.

LA.3.RI.1 Identify the central idea and explain how key details support that idea.

LA.3.RI.2 Explain the relationship between individuals, historical events, scientific ideas or concepts, or steps in a process.

Author's Craft | Citing relevant and thorough evidence to support ideas, evaluate the development and interaction of individuals, ideas, and events in grade-level informational text.

LA.3.RI.3 Determine and explain the author's purpose in an informational text.

LA.3.RI.4 Explain how text features (titles, headings, table of contents, glossaries, captions, graphs, maps, and/or other visuals) contribute to meaning.

Knowledge and Ideas | Citing relevant and thorough textual evidence to support ideas, evaluate how an author's perspective or use of point of view shapes the style and meaning of grade-level informational text.

LA.3.RI.5 Compare and contrast the two most important ideas and key details presented by multiple informational texts on the same topic.

LA.3.RI.6 Identify an author's claim(s) and explain how the author supports the claim(s) in the text.

LA.3.RI.7 Compare and contrast topics and/or patterns of events in a range of informational texts.

Do not select the following standard - locally assessed only:

Range of Reading and Level of Text Complexity | Read and comprehend complex, grade-level informational text independently and proficiently.

LA.3.RI.8 Read and comprehend a wide range of informational texts of appropriate complexity for Grade 3 independently and proficiently.

Instructional Considerations

- A claim refers to an author's primary argument and is supported by textual evidence.
- Author's craft refers to the techniques an author uses to develop and support a claim.
- Point of view refers to the vantage point from which a story is told, while perspective is an author's attitude or belief that is based on personal knowledge and/or experiences.

Page 48

Approved by the Nebraska State Board of Education on September 2, 2021



VOCABULARY

Acquisition and Use | Build and use a range of conversational, academic, and discipline-specific grade-level vocabulary and apply to reading, writing, speaking, and listening.

LA.3.V.1 Acquire and use grade-level academic vocabulary appropriately.

- a. Use sentence-level context clues to determine the meaning of a word or phrase.
- b. Use affixes to determine the meaning of unknown words (e.g., comfortable, uncomfortable).
- c. Use known root words to determine the meaning of unknown words (e.g., company, companion).

Do not select the following sub-standard - locally assessed only:

- d. Determine the meanings of key words and phrases using reference materials and classroom resources.

Context and Connotation | Determine or clarify the meaning of unknown and multiple-meaning words and phrases, choosing flexibly from a range of strategies.

LA.3.V.2 Interpret an author's use of figurative, connotative, and technical language in grade-level literary and informational text.

- a. Distinguish between literal and nonliteral meanings of words and phrases in context (e.g., take steps).
- b. Identify real-life connections between words and their use (e.g., describe people who are friendly or helpful).
- c. Distinguish nuances of meaning between related words that describe states of mind or degrees of certainty (e.g., believed, suspected).

Instructional Considerations

- Academic vocabulary refers to words likely to appear in a variety of content area texts, at or above grade-level, and typically requires explicit instruction. Students should be encouraged to use newly acquired terms frequently in speaking and writing.
- The vast majority of academic vocabulary should derive from complex texts—a careful review of texts for challenging words that are central to understanding the meaning of the text, including figurative language, should determine which vocabulary is taught explicitly (sometimes in advance of reading).
- Include a word study component that includes prefixes, root words, and suffixes to accompany text-based methods of vocabulary development.
- Reading aloud to students using texts that are two grade levels higher than their reading level is an evidence-based practice for activating prior knowledge and building vocabulary.

Page 49

Approved by the Nebraska State Board of Education on September 2, 2021



WRITING

Production of Writing | Use a recursive writing process to produce clear and coherent writing appropriate to the discipline, audience, and/or context.

LA.3.W.1 Write paragraphs using a variety of sentence types.

- Capitalize proper nouns (e.g., historic periods, nationalities, languages), proper adjectives (e.g., South American), and appropriate words in titles.
- Use commas in addresses and commas and quotation marks in dialogue; use an apostrophe to form and use possessives.
- Use frequently occurring nouns (e.g., concrete and abstract), verbs (regular and irregular), and simple verb tenses.
- Distinguish between and use coordinating and subordinating conjunctions and independent and dependent clauses.
- Explain the function of adjectives and adverbs in simple, compound, and complex sentences.
- Use correct subject-verb and pronoun-antecedent agreement in speaking and writing.
- Use frequently occurring prepositions and prepositional phrases.

Do not select the following standard/sub-standards - locally assessed only:

LA.3.W.2 Use a recursive writing process to develop, strengthen, and produce writing appropriate to the audience, purpose, and discipline.

- Use prewriting activities and resources to plan, organize, and draft writing.
- Adapt writing processes to sustain engagement in short and long-term writing tasks of increasing length and complexity.
- Improve and clarify the content, structure, and organization of writing by revising, considering feedback from adults and peers.
- Improve and clarify writing by editing and proofreading, considering feedback from adults and peers.
- Use or decipher multiple formats of print and digital text (e.g., manuscript, cursive, font, graphics, symbols).
- Use appropriate print and digital/multimedia tools to produce, enhance, and/or publish writing individually or in collaboration with peers.

Page 50

Approved by the Nebraska State Board of Education on September 2, 2021



Modes of Writing | Write in a variety of modes for a variety of purposes and audiences across disciplines.

LA.3.W.3 Write creative and/or expressive pieces that describe a well-developed event or experience.

- a. Engage and orient the reader by establishing a situation and introducing a narrator and/or character(s).
- b. Include descriptive details about characters, events, or settings.
- c. Use words and phrases to signal a sequence of events.
- d. Provide a closure related to the creative or expressive event or experience.

LA.3.W.4 Write opinion pieces with supporting reasons and/or evidence.

- a. Introduce a topic or text, state an opinion, and develop a structure that includes reasons and/or evidence.
- b. Use linking words and phrases to connect opinions and reasons.
- c. Provide a concluding statement or section related to the opinion.

LA.3.W.5 Write informative/explanatory pieces to examine a topic or text and convey ideas and information.

- a. Introduce a topic and group related information together, including illustrations when useful to provide clarity.
- b. Develop the topic with information (e.g., facts, definitions, details) clearly related to the topic.
- c. Use linking words and phrases and key vocabulary to connect ideas and categories of information.
- d. Provide a concluding statement or section related to the topic.

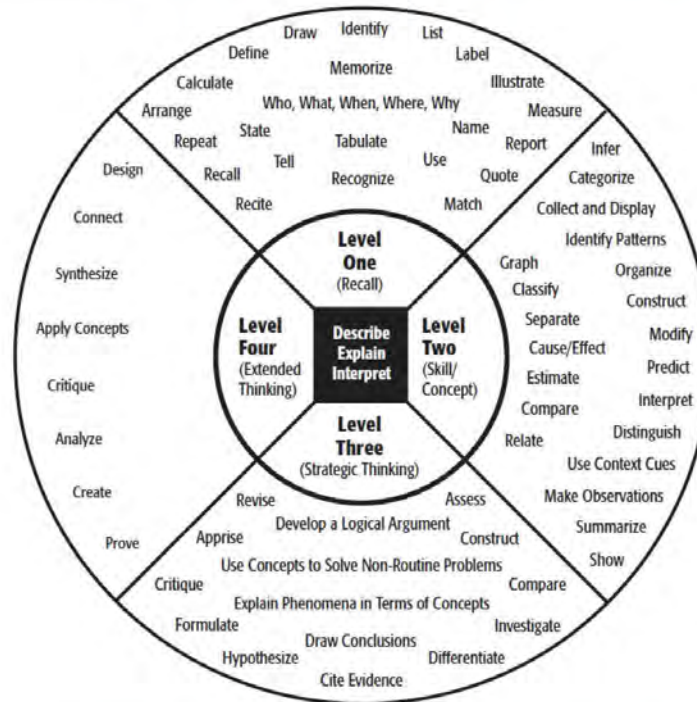
LA.3.W.6 Locate evidence from literary and/or informational text sources to answer questions about a topic.

- a. Paraphrase information from sources to support ideas while avoiding plagiarism.
Do not select the following sub-standard:
- b. Identify print and digital tools to gather information and ideas to answer questions.
- c. Sort evidence into categories using an appropriate note-taking format to collect and organize information.
Do not select the following sub-standards:
- d. Demonstrate academic integrity by avoiding overreliance on any one source and referencing sources in writing and speaking; provide a list of sources.
- e. Practice safe and ethical behavior when communicating and interacting with others digitally (e.g., safe information to share, utilize appropriate sites and materials, appropriate language use, respect diverse perspectives).



Appendix F. Cognitive Complexity (DOK Wheel)

Depth of Knowledge (DOK) Levels



Level One Activities	Level Two Activities	Level Three Activities	Level Four Activities
Recall elements and details of story structure, such as sequence of events, character, plot and setting. Conduct basic mathematical calculations. Label locations on a map. Represent in words or diagrams a scientific concept or relationship. Perform routine procedures like measuring length or using punctuation marks correctly. Describe the features of a place or people.	Identify and summarize the major events in a narrative. Use context cues to identify the meaning of unfamiliar words. Solve routine multiple-step problems. Describe the cause/effect of a particular event. Identify patterns in events or behavior. Formulate a routine problem given data and conditions. Organize, represent and interpret data.	Support ideas with details and examples. Use voice appropriate to the purpose and audience. Identify research questions and design investigations for a scientific problem. Develop a scientific model for a complex situation. Determine the author's purpose and describe how it affects the interpretation of a reading selection. Apply a concept in other contexts.	Conduct a project that requires specifying a problem, designing and conducting an experiment, analyzing its data, and reporting results/ solutions. Apply mathematical model to illuminate a problem or situation. Analyze and synthesize information from multiple sources. Describe and illustrate how common themes are found across texts from different cultures. Design a mathematical model to inform and solve a practical or abstract situation.

Webb, Norman L. and others. "Web Alignment Tool" 24 July 2005. Wisconsin Center of Educational Research. University of Wisconsin-Madison. 2 Feb. 2006. <<http://www.wcer.wisc.edu/WAT/index.asp>>.



Appendix G. Achievement Level Descriptors (Grade 3 Example)

Indicator No.	Indicator Text	Developing	On Track	Advanced
		With a range of texts with text complexity commonly found in Grade 3, a student performing in Developing can likely	With a range of texts with text complexity commonly found in Grade 3, a student performing in On Track can likely	With a range of texts with text complexity commonly found at the intersection of Grade 3 and Grade 4, a student performing in Advanced can likely
Reading Prose and Poetry				
LA.3.RP.1	Identify the central message or lesson in a literary text and explain how key details support that idea.	Identify the central message or lesson in a literary text.	Identify the central message or lesson in a literary text and explain how key details support that idea.	Analyze the central message or lesson in a literary text and explain how key details support that idea.
LA.3.RP.2	Explain how characters respond to major events and challenges in a literary text.	Identify the major events and/or challenges that characters face in a literary text.	Explain how characters respond to major events and challenges in a literary text.	Analyze how characters respond to major events and challenges in a literary text, drawing on specific details such as a character's thoughts, words, or actions.



Appendix H. Correlation Analysis

The correlation between Depth of Knowledge and Achievement Level Descriptors was examined across grade levels. The results revealed a moderate correlation in Grades 4-8, with values ranging from $r = .31$ to $r = .45$. In Grade 3, a stronger correlation of $r = .57$ was observed (Table H1).

Table H1. Correlation between DOK and ALD by Grade

Grade	Correlation
ELA 3	.57
ELA 4	.45
ELA 5	.39
ELA 6	.37
ELA 7	.31
ELA 8	.41



Appendix I. Number of Unique and Shared Items by Grade, Test Event, and Strand

The number of unique and shared items by grade, test event, and strand were examined. Results showed that many writing items were shared across test events and achievement levels, except for Grade 4, which showed a greater number of unique writing items. These results may indicate that the pool of writing items is shallow and/or the CAT test algorithm is not selecting unique writing items across test events and achievement levels.



Table 11. Number of Unique and Shared Items by Test Event and Strand – Grade 3

Achievement Level	Test Event	Number of Items	RP		RI		V		W	
			Unique	Shared	Unique	Shared	Unique	Shared	Unique	Shared
Developing	1	30	4	4	9	3	4	1	0	6
	2	32	5	2	10	0	7	2	1	5
	3	32	5	2	5	6	6	2	0	6
	4	31	8	0	7	2	7	1	0	6
On Track	1	28	3	6	2	4	1	5	0	6
	2	28	5	4	4	3	3	2	0	6
	3	31	3	5	7	3	5	3	0	6
	4	27	2	5	6	2	1	4	0	6
Advanced	1	31	9	0	8	4	4	0	0	6
	2	28	7	2	4	4	3	2	2	4
	3	29	0	8	3	4	4	3	0	6
	4	28	1	9	5	0	5	1	0	6

Table 12. Number of Unique and Shared Items by Test Event and Strand – Grade 4

Achievement Level	Test Event	Number of Items	RP		RI		V		W	
			Unique	Shared	Unique	Shared	Unique	Shared	Unique	Shared
Developing	1	28	5	3	4	4	1	5	4	2
	2	24	4	3	0	4	4	2	2	4
	3	29	7	3	4	2	3	4	4	2
	4	27	5	2	6	0	5	0	3	3
On Track	1	23	4	3	1	6	2	1	4	2
	2	28	4	5	1	5	3	3	3	3
	3	27	5	2	5	3	3	2	1	5
	4	27	9	0	4	3	3	1	2	4
Advanced	1	27	5	3	5	3	1	4	2	4
	2	27	1	6	4	6	3	1	2	4
	3	27	7	0	2	7	4	2	0	6
	4	29	6	2	4	3	5	3	0	6



Table I3. Number of Unique and Shared Items by Test Event and Strand – Grade 5

Achievement Level	Test Event	Number of Items	RP		RI		V		W	
			Unique	Shared	Unique	Shared	Unique	Shared	Unique	Shared
Developing	1	32	5	5	10	0	5	1	5	1
	2	25	4	5	2	2	1	5	1	5
	3	29	5	2	9	1	6	0	0	6
	4	32	5	6	6	2	6	1	2	4
On Track	1	26	1	6	3	5	0	5	0	6
	2	29	3	5	6	3	4	3	1	5
	3	30	5	2	5	3	6	3	0	6
	4	22	3	4	4	2	1	1	1	5
Advanced	1	29	7	4	2	5	0	5	0	6
	2	30	4	3	7	3	5	2	0	6
	3	23	0	5	7	0	3	1	0	6
	4	24	0	1	5	5	3	4	0	6

Table I4. Number of Unique and Shared Items by Test Event and Strand – Grade 6

Achievement Level	Test Event	Number of Items	RP		RI		V		W	
			Unique	Shared	Unique	Shared	Unique	Shared	Unique	Shared
Developing	1	23	4	6	3	0	2	2	3	3
	2	28	7	0	6	3	2	3	4	2
	3	29	7	0	5	4	4	3	4	2
	4	29	6	2	2	7	1	5	1	5
On Track	1	28	6	1	1	5	6	2	1	5
	2	30	3	7	4	3	2	5	0	6
	3	27	1	7	3	3	4	3	1	5
	4	28	4	4	2	6	2	4	1	5
Advanced	1	29	1	6	4	5	1	6	0	6
	2	29	2	5	4	5	1	6	0	6
	3	26	0	8	2	5	0	5	0	6
	4	28	1	7	4	5	3	2	0	6



Table 15. Number of Unique and Shared Items by Test Event and Strand – Grade 7

Achievement Level	Test Event	Number of Items	RP		RI		V		W	
			Unique	Shared	Unique	Shared	Unique	Shared	Unique	Shared
Developing	1	31	5	2	10	1	4	3	1	4
	2	29	4	3	9	0	4	3	2	4
	3	28	0	7	8	2	6	0	3	2
	4	31	8	0	6	2	6	3	1	4
On Track	1	23	0	4	5	4	3	1	2	4
	2	27	4	6	2	5	2	3	1	5
	3	28	2	5	4	6	4	1	2	4
	4	27	5	3	5	3	1	4	0	6
Advanced	1	27	1	6	7	2	3	2	0	6
	2	23	0	5	1	6	2	3	0	4
	3	27	2	6	4	5	3	2	0	6
	4	27	0	7	5	4	3	2	1	4

Table 16. Number of Unique and Shared Items by Test Event and Strand – Grade 8

Achievement Level	Test Event	Number of Items	RP		RI		V		W	
			Unique	Shared	Unique	Shared	Unique	Shared	Unique	Shared
Developing	1	29	10	1	6	0	6	1	3	3
	2	30	8	3	9	0	2	2	7	0
	3	24	0	4	1	4	2	4	2	4
	4	32	3	4	9	0	8	2	0	6
On Track	1	28	7	2	0	7	5	0	0	6
	2	22	2	5	0	3	1	4	0	6
	3	29	1	6	6	3	3	4	2	4
	4	30	6	1	8	0	7	1	0	6
Advanced	1	28	2	6	3	4	5	1	0	6
	2	27	1	6	3	6	0	5	0	6
	3	25	1	3	9	0	5	1	0	6
	4	27	2	6	4	3	3	1	0	6



Appendix J. DOK by Grade, Test Event, and Strand

Table J1. DOK by Grade, Test Event, and Strand – Grade 3

Achievement Level	Test Event	Strand	Number of Items	DOK 1	DOK 2	DOK 3
Developing	1	RP	8	0 (0%)	6 (75%)	2 (25%)
		RI	11	1 (9%)	9 (82%)*	1 (9%)
		V	5	0 (0%)	4 (80%)	1 (20%)
		W	6	0 (0%)	6 (100%)	0 (0%)
	2	RP	7	4 (57%)	0 (0%)	3 (43%)
		RI	10	0 (0%)	8 (80%)	2 (20%)
		V	9	1 (11%)	7 (78%)	1 (11%)
		W	6	0 (0%)	6 (100%)	0 (0%)
	3	RP	7	0 (0%)	4 (57%)	3 (43%)
		RI	11	1 (9%)	7 (64%)	3 (27%)
		V	8	0 (0%)	7 (88%)	1 (13%)
		W	6	0 (0%)	6 (100%)	0 (0%)
	4	RP	8	1 (13%)	3 (38%)	4 (50%)
		RI	9	1 (11%)	3 (33%)	5 (56%)
		V	8	1 (13%)	5 (63%)	2 (25%)
		W	6	0 (0%)	6 (100%)	0 (0%)
On Track	1	RP	9	1 (11%)	3 (33%)	5 (56%)
		RI	6	0 (0%)	3 (50%)	3 (50%)
		V	6	0 (0%)	5 (83%)	1 (17%)
		W	6	0 (0%)	6 (100%)	0 (0%)
	2	RP	9	0 (0%)	6 (67%)	3 (33%)
		RI	7	0 (0%)	6 (86%)	1 (14%)
		V	5	0 (0%)	4 (80%)	1 (20%)
		W	6	0 (0%)	6 (100%)	0 (0%)
	3	RP	8	1 (13%)	3 (38%)	4 (50%)
		RI	9	1 (11%)	7 (78%)*	1 (11%)
		V	8	0 (0%)	8 (100%)	0 (0%)
		W	6	0 (0%)	6 (100%)	0 (0%)
	4	RP	7	1 (14%)	4 (57%)	2 (29%)
		RI	8	0 (0%)	8 (100%)	0 (0%)
		V	5	0 (0%)	3 (60%)	2 (40%)
		W	6	0 (0%)	6 (100%)	0 (0%)



Achievement Level	Test Event	Strand	Number of Items	DOK 1	DOK 2	DOK 3
Advanced	1	RP	9	0 (0%)	6 (67%)	3 (33%)
		RI	12	3 (25%)	7 (58%)	2 (17%)
		V	4	0 (0%)	2 (50%)	2 (50%)
		W	6	0 (0%)	6 (100%)	0 (0%)
	2	RP	9	1 (11%)	6 (67%)	2 (22%)
		RI	8	0 (0%)	5 (63%)	3 (38%)
		V	5	0 (0%)	3 (60%)	2 (40%)
		W	6	0 (0%)	5 (83%)	1 (17%)
	3	RP	8	0 (0%)	5 (63%)	3 (38%)
		RI	7	0 (0%)	5 (71%)	2 (29%)
		V	7	0 (0%)	6 (86%)	1 (14%)
		W	6	0 (0%)	6 (100%)	0 (0%)
	4	RP	10	0 (0%)	7 (70%)	3 (30%)
		RI	5	1 (20%)	4 (80%)	0 (0%)
		V	6	0 (0%)	5 (83%)	1 (17%)
		W	6	0 (0%)	6 (100%)	0 (0%)

Notes.

- One test item in the Developing 1 and On Track 3 test events was split between RI and RI; it received a DOK of 2.
- For the On Track achievement level, test events #1, #2, and #4 had one item rated as "None" for the standard; therefore, this item was not included in this table.
- For the Advanced achievement level, test events #3 and #4 had one item rated as "None" for the standard; therefore, this item was not included in this table.



Table J2. DOK by Grade, Test Event, and Strand – Grade 4

Achievement Level	Test Event	Strand	Number of Items	DOK 1	DOK 2	DOK 3
Developing	1	RP	8	2 (25%)	5 (63%)	1 (13%)
		RI	8	5 (63%)	3 (38%)	0 (0%)
		V	6	3 (50%)	3 (50%)	0 (0%)
		W	6	2 (33%)	4 (67%)	0 (0%)
	2	RP	7	2 (29%)	5 (71%)	0 (0%)
		RI	4	4 (100%)	0 (0%)	0 (0%)
		V	6	3 (50%)	3 (50%)	0 (0%)
		W	6	0 (0%)	5 (83%)	1 (17%)
	3	RP	10	2 (20%)	5 (50%)	3 (30%)
		RI	6	2 (33%)	4 (67%)	0 (0%)
		V	7	4 (57%)	3 (43%)	0 (0%)
		W	6	2 (33%)	4 (67%)	0 (0%)
	4	RP	7	1 (14%)	3 (43%)	3 (43%)
		RI	6	4 (67%)	1 (17%)	1 (17%)
		V	5	0 (0%)	5 (100%)	0 (0%)
		W	6	0 (0%)	5 (83%)	1 (17%)
On Track	1	RP	7	2 (29%)	3 (43%)	2 (29%)
		RI	7	3 (43%)	4 (57%)	0 (0%)
		V	3	0 (0%)	3 (100%)	0 (0%)
		W	6	1 (17%)	4 (67%)	1 (17%)
	2	RP	9	0 (0%)	9 (100%)	0 (0%)
		RI	6	1 (17%)	5 (83%)	0 (0%)
		V	6	4 (67%)	2 (33%)	0 (0%)
		W	6	1 (17%)	5 (83%)	0 (0%)
	3	RP	7	0 (0%)	5 (71%)	2 (29%)
		RI	8	2 (25%)	5 (63%)	1 (13%)
		V	5	1 (20%)	4 (80%)	0 (0%)
		W	6	1 (17%)	4 (67%)	1 (17%)
	4	RP	9	1 (11%)	7 (78%)	1 (11%)
		RI	7	1 (14%)	5 (71%)	1 (14%)
		V	4	2 (50%)	2 (50%)	0 (0%)
		W	6	1 (17%)	5 (83%)	0 (0%)



Achievement Level	Test Event	Strand	Number of Items	DOK 1	DOK 2	DOK 3
Advanced	1	RP	8	0 (0%)	7 (88%)	1 (13%)
		RI	8	2 (25%)	5 (63%)	1 (13%)
		V	5	1 (20%)	4 (80%)	0 (0%)
		W	6	0 (0%)	6 (100%)	0 (0%)
	2	RP	7	0 (0%)	5 (71%)	2 (29%)
		RI	9	1 (11%)	6 (66%)	2 (22%)*
		V	4	2 (50%)	2 (50%)	0 (0%)
		W	6	0 (0%)	5 (83%)	1 (17%)
	3	RP	7	0 (0%)	4 (57%)	3 (43%)
		RI	8	2 (25%)	3 (38%)	3 (38%)*
		V	6	1 (17%)	5 (83%)	0 (0%)
		W	6	0 (0%)	5 (83%)	1 (17%)
	4	RP	8	2 (25%)	5 (63%)	1 (13%)
		RI	7	1 (14%)	5 (71%)	1 (14%)
		V	8	4 (50%)	4 (50%)	0 (0%)
		W	6	0 (0%)	5 (83%)	1 (17%)

Notes.

- One test item in the Advanced 2 and Advanced 3 test events was split between RI and RI; it received a DOK of 3.
- For the Developing achievement level, test event #2 had one item rated as "None" for the standard; therefore, this item was not included in this table.
- For the Developing achievement level, test event #4 had three items rated as "None" for the standard; therefore, this item was not included in this table.
- For the On Track achievement level, test events #2, #3, and #4 had one item rated as "None" for the standard; therefore, this item was not included in this table.
- For the Advanced achievement level, test event #2 had one item rated as "None" for the standard; therefore, this item was not included in this table.



Table J3. DOK by Grade, Test Event, and Strand – Grade 5

Achievement Level	Test Event	Strand	Number of Items	DOK 1	DOK 2	DOK 3
Developing	1	RP	10	0 (0%)	10 (100%)	0 (0%)
		RI	10	1 (10%)	9 (90%)	0 (0%)
		V	6	1 (17%)	5 (83%)	0 (0%)
		W	6	1 (17%)	5 (83%)	0 (0%)
	2	RP	9	0 (0%)	8 (89%)	1 (11%)
		RI	4	0 (0%)	3 (75%)	1 (25%)
		V	6	2 (33%)	4 (67%)	0 (0%)
		W	6	0 (0%)	6 (100%)	0 (0%)
	3	RP	7	1 (14%)	6 (86%)	0 (0%)
		RI	10	1 (10%)	7 (70%)	2 (20%)
		V	6	1 (17%)	5 (83%)	0 (0%)
		W	6	0 (0%)	6 (100%)	0 (0%)
	4	RP	11	2 (18%)	9 (82%)	0 (0%)
		RI	8	0 (0%)	8 (100%)	0 (0%)
		V	7	3 (43%)	4 (57%)	0 (0%)
		W	6	0 (0%)	6 (100%)	0 (0%)
On Track	1	RP	7	0 (0%)	7 (100%)	0 (0%)
		RI	8	0 (0%)	7 (88%)	1 (13%)
		V	5	2 (40%)	3 (60%)	0 (0%)
		W	6	0 (0%)	6 (100%)	0 (0%)
	2	RP	7	0 (0%)	6 (86%)	1 (14%)
		RI	9	0 (0%)	9 (100%)	0 (0%)
		V	7	2 (29%)	5 (71%)	0 (0%)
		W	6	0 (0%)	6 (100%)	0 (0%)
	3	RP	7	0 (0%)	6 (86%)	1 (14%)
		RI	8	1 (13%)	7 (88%)	0 (0%)
		V	9	2 (22%)	7 (78%)	0 (0%)
		W	6	0 (0%)	6 (100%)	0 (0%)
	4	RP	7	0 (0%)	6 (86%)	1 (14%)
		RI	6	1 (17%)	5 (83%)	0 (0%)
		V	2	0 (0%)	2 (100%)	0 (0%)
		W	6	0 (0%)	6 (100%)	0 (0%)



Achievement Level	Test Event	Strand	Number of Items	DOK 1	DOK 2	DOK 3
Advanced	1	RP	11	0 (0%)	11 (100%)	0 (0%)
		RI	7	0 (0%)	6 (86%)	1 (14%)
		V	5	2 (40%)	3 (60%)	0 (0%)
		W	6	0 (0%)	6 (100%)	0 (0%)
	2	RP	7	0 (0%)	7 (100%)	0 (0%)
		RI	10	1 (10%)	7 (70%)	2 (20%)
		V	7	1 (14%)	6 (87%)	0 (0%)
		W	6	0 (0%)	6 (100%)	0 (0%)
	3	RP	5	0 (0%)	4 (80%)	1 (20%)
		RI	7	1 (14%)	6 (86%)	0 (0%)
		V	4	0 (0%)	4 (100%)	0 (0%)
		W	6	0 (0%)	6 (100%)	0 (0%)
	4	RP	1	0 (0%)	1 (100%)	0 (0%)
		RI	10	0 (0%)	9 (90%)	1 (10%)
		V	7	2 (29%)	5 (71%)	0 (0%)
		W	6	0 (0%)	6 (100%)	0 (0%)

Notes.

- One test item in the Developing 1 test event was split between RP and RP; it received a DOK of 2.
- One test item in the Developing 2 test event was split between W and W; it received a DOK of 2.
- One test item in the On Track 2 test event was split between RP and RP; it received a DOK of 2.
- For the On Track achievement level, test event #4 had one item rated as "None" for the standard; therefore, this item was not included in this table.
- For the Advanced achievement level, test event #3 had one item rated as "None" for the standard; therefore, this item was not included in this table.



Table J4. DOK by Grade, Test Event, and Strand – Grade 6

Achievement level	Test Event	Strand	Number of Items	DOK 1	DOK 2	DOK 3
Developing	1	RP	10	1 (10%)	8 (80%)	1 (10%)
		RI	3	0 (0%)	3 (100%)	0 (0%)
		V	4	0 (0%)	3 (75%)	1 (25%)
		W	6	0 (0%)	4 (67%)	2 (33%)
	2	RP	7	2 (29%)	3 (43%)	2 (29%)
		RI	9	3 (33%)	5 (56%)	1 (11%)
		V	5	2 (40%)	3 (60%)	0 (0%)
		W	6	0 (0%)	5 (83%)	1 (17%)
	3	RP	7	2 (29%)	5 (71%)	0 (0%)
		RI	9	2 (22%)	5 (56%)	2 (22%)
		V	7	2 (29%)	5 (71%)	0 (0%)
		W	6	0 (0%)	3 (50%)	3 (50%)
	4	RP	8	0 (0%)	8 (100%)	0 (0%)
		RI	9	2 (22%)	6 (67%)	1 (11%)
		V	6	2 (33%)	4 (67%)	0 (0%)
		W	6	0 (0%)	4 (67%)	2 (33%)
On Track	1	RP	7	0 (0%)	6 (86%)	1 (14%)
		RI	6	0 (0%)	5 (83%)	1 (17%)
		V	8	0 (0%)	7 (88%)	1 (13%)
		W	6	0 (0%)	4 (67%)	2 (33%)
	2	RP	10	2 (20%)	8 (80%)	0 (0%)
		RI	7	1 (14%)	6 (86%)	0 (0%)
		V	7	2 (29%)	5 (71%)	0 (0%)
		W	6	0 (0%)	3 (50%)	3 (50%)
	3	RP	8	0 (0%)	7 (88%)	1 (13%)
		RI	6	0 (0%)	6 (100%)	0 (0%)
		V	7	0 (0%)	7 (100%)	0 (0%)
		W	6	0 (0%)	3 (50%)	3 (50%)
	4	RP	8	0 (0%)	4 (50%)	4 (50%)
		RI	8	1 (13%)	6 (75%)	1 (13%)
		V	6	0 (0%)	5 (83%)	1 (17%)
		W	6	0 (0%)	3 (50%)	3 (50%)



Achievement level	Test Event	Strand	Number of Items	DOK 1	DOK 2	DOK 3
Advanced	1	RP	7	0 (0%)	6 (86%)	1 (14%)
		RI	9	0 (0%)	8 (89%)	1 (11%)
		V	7	0 (0%)	7 (100%)	0 (0%)
		W	6	0 (0%)	1 (17%)	5 (83%)
	2	RP	7	0 (0%)	5 (71%)	2 (29%)
		RI	9	0 (0%)	8 (89%)	1 (11%)
		V	7	0 (0%)	7 (100%)	0 (0%)
		W	6	0 (0%)	1 (17%)	5 (83%)
	3	RP	8	0 (0%)	7 (88%)	1 (13%)
		RI	7	0 (0%)	6 (86%)	1 (14%)
		V	5	0 (0%)	5 (100%)	0 (0%)
		W	6	0 (0%)	2 (33%)	4 (67%)
	4	RP	8	1 (13%)	7 (88%)	0 (0%)
		RI	9	3 (33%)	5 (56%)	1 (11%)
		V	5	1 (20%)	4 (80%)	0 (0%)
		W	6	0 (0%)	1 (17%)	5 (83%)

Notes.

- For the Developing achievement level, test event #2 had one item rated as "None" for the standard; therefore, this item was not included in this table.
- For the On Track achievement level, test event #1 had one item rated as "None" for the standard; therefore, this item was not included in this table.



Table J5. DOK by Grade, Test Event, and Strand – Grade 7

Achievement level	Test Event	Strand	Number of Items	DOK 1	DOK 2	DOK 3
Developing	1	RP	7	1 (14%)	4 (57%)	2 (29%)
		RI	11	3 (27%)	7 (64%)	1 (9%)
		V	7	0 (0%)	6 (86%)	1 (14%)
		W	5	0 (0%)	3 (60%)	1 (20%)
	2	RP	7	0 (0%)	5 (71%)	2 (29%)
		RI	9	0 (0%)	7 (78%)	2 (22%)
		V	7	2 (29%)	5 (71%)	0 (0%)
		W	6	0 (0%)	4 (67%)	2 (33%)
	3	RP	7	0 (0%)	4 (57%)	3 (43%)
		RI	9	1 (11%)	5 (56%)	2 (22%)
		V	6	0 (0%)	5 (83%)	1 (17%)
		W	5	0 (0%)	4 (80%)	1 (20%)
	4	RP	8	1 (13%)	4 (50%)	3 (38%)
		RI	8	1 (13%)	5 (63%)	2 (25%)
		V	9	0 (0%)	8 (89%)	1 (11%)
		W	5	0 (0%)	3 (60%)	2 (40%)
On Track	1	RP	4	0 (0%)	1 (25%)	3 (75%)
		RI	9	2 (22%)	3 (33%)	4 (44%)
		V	4	0 (0%)	4 (100%)	0 (0%)
		W	6	0 (0%)	3 (50%)	2 (33%)
	2	RP	10	0 (0%)	7 (70%)	3 (30%)
		RI	7	0 (0%)	6 (86%)	1 (14%)
		V	5	0 (0%)	4 (80%)	1 (20%)
		W	6	0 (0%)	3 (50%)	3 (50%)
	3	RP	7	0 (0%)	5 (71%)	2 (29%)
		RI	10	2 (20%)	8 (80%)	0 (0%)
		V	5	0 (0%)	5 (100%)	0 (0%)
		W	6	0 (0%)	4 (67%)	2 (33%)
	4	RP	8	0 (0%)	7 (88%)	1 (13%)
		RI	8	0 (0%)	4 (50%)	4 (50%)
		V	5	0 (0%)	4 (80%)	1 (20%)
		W	6	0 (0%)	3 (50%)	3 (50%)



Achievement level	Test Event	Strand	Number of Items	DOK 1	DOK 2	DOK 3
Advanced	1	RP	7	0 (0%)	5 (71%)	2 (29%)
		RI	9	1 (11%)	5 (56%)	3 (33%)
		V	5	0 (0%)	5 (100%)	0 (0%)
		W	6	0 (0%)	3 (50%)	3 (50%)
	2	RP	5	0 (0%)	5 (100%)	0 (0%)
		RI	7	0 (0%)	5 (71%)	2 (29%)
		V	5	0 (0%)	4 (80%)	1 (20%)
		W	4	0 (0%)	3 (75%)	1 (25%)
	3	RP	7	0 (0%)	4 (57%)	3 (43%)
		RI	9	0 (0%)	8 (89%)	1 (11%)
		V	5	0 (0%)	1 (20%)	4 (80%)
		W	6	0 (0%)	4 (67%)	2 (33%)
	4	RP	7	0 (0%)	7 (100%)	0 (0%)
		RI	9	0 (0%)	5 (56%)	4 (44%)
		V	5	0 (0%)	4 (80%)	1 (20%)
		W	5	0 (0%)	3 (60%)	2 (40%)

Notes.

- One test item in the Developing 3 test event was split between RI and RI; it did not receive a DOK rating.
- One test item in the On Track 2 test event was split between RP and V; it received a DOK of 2.
- One test item in the Advanced 3 test event was split between RP and RP; it received a DOK of 3.
- For the Developing achievement level, test events #1, #3, and #4 had one item rated as "None" for the standard; therefore, this item was not included in this table.
- For the Advanced achievement level, test event #2 had two items rated as "None" for the standard; therefore, this item was not included in this table.
- For the Advanced achievement level, test event #4 had one item rated as "None" for the standard; therefore, this item was not included in this table.



Table J6. DOK by Grade, Test Event, and Strand – Grade 8

Achievement level	Test Event	Strand	Number of Items	DOK 1	DOK 2	DOK 3
Developing	1	RP	10	0 (0%)	10 (100%)	0 (0%)
		RI	6	1 (17%)	5 (83%)	0 (0%)
		V	7	1 (14%)	5 (71%)	1 (14%)
		W	6	0 (0%)	6 (100%)	0 (0%)
	2	RP	11	2 (18%)	9 (82%)	0 (0%)
		RI	9	2 (22%)	7 (78%)	0 (0%)
		V	4	1 (25%)	3 (75%)	0 (0%)
		W	6	0 (0%)	6 (100%)	0 (0%)
	3	RP	4	0 (0%)	3 (75%)	1 (25%)
		RI	5	1 (20%)	2 (40%)	2 (40%)
		V	6	1 (17%)	5 (83%)	0 (0%)
		W	6	0 (0%)	5 (83%)	1 (17%)
	4	RP	7	1 (14%)	5 (71%)	1 (14%)
		RI	9	4 (44%)	4 (44%)	1 (11%)
		V	10	1 (10%)	9 (90%)	0 (0%)
		W	6	0 (0%)	5 (83%)	1 (17%)
On Track	1	RP	9	0 (0%)	9 (100%)	0 (0%)
		RI	7	1 (14%)	6 (86%)	0 (0%)
		V	5	1 (20%)	4 (80%)	0 (0%)
		W	6	0 (0%)	6 (100%)	0 (0%)
	2	RP	7	0 (0%)	6 (86%)	1 (14%)
		RI	3	0 (0%)	1 (33%)	2 (67%)
		V	5	0 (0%)	5 (100%)	0 (0%)
		W	6	0 (0%)	6 (100%)	0 (0%)
	3	RP	7	1 (14%)	4 (57%)	2 (29%)
		RI	9	1 (11%)	7 (78%)	1 (11%)
		V	7	1 (14%)	6 (86%)	0 (0%)
		W	6	0 (0%)	6 (100%)	0 (0%)
	4	RP	7	0 (0%)	6 (86%)	1 (14%)
		RI	8	0 (0%)	7 (88%)	1 (13%)
		V	8	2 (25%)	6 (75%)	0 (0%)
		W	6	0 (0%)	5 (83%)	1 (17%)



Achievement level	Test Event	Strand	Number of Items	DOK 1	DOK 2	DOK 3
Advanced	1	RP	8	1 (13%)	6 (75%)	1 (13%)
		RI	7	2 (29%)	4 (57%)	1 (14%)
		V	6	0 (0%)	6 (100%)	0 (0%)
		W	6	0 (0%)	6 (100%)	0 (0%)
	2	RP	7	1 (14%)	5 (71%)	1 (14%)
		RI	9	0 (0%)	7 (78%)	2 (22%)
		V	5	1 (20%)	4 (80%)	0 (0%)
		W	6	0 (0%)	6 (100%)	0 (0%)
	3	RP	4	0 (0%)	2 (50%)	2 (50%)
		RI	9	0 (0%)	9 (100%)	0 (0%)
		V	6	1 (17%)	5 (83%)	0 (0%)
		W	6	0 (0%)	6 (100%)	0 (0%)
	4	RP	8	1 (13%)	6 (75%)	1 (13%)
		RI	7	4 (57%)	3 (43%)	0 (0%)
		V	4	0 (0%)	4 (100%)	0 (0%)
		W	6	0 (0%)	6 (100%)	0 (0%)

Notes.

- For the Developing achievement level, test event #3 had three items rated as "None" for the standard; therefore, this item was not included in this table.
- For the On Track achievement level, test events #1, #2, and #4 had one item rated as "None" for the standard; therefore, this item was not included in this table.
- For the Advanced achievement level, test event #1 had one item rated as "None" for the standard; therefore, this item was not included in this table.
- For the Advanced achievement level, test event #4 had two items rated as "None" for the standard; therefore, this item was not included in this table.
- One test item in the Developing 1 test event was split between RP and RP; it received a DOK of 2.
- One test item in the Developing 2 test event was split between W and W; it received a DOK of 2.



Appendix K. ALD by Grade, Test Event, and Strand

Table K1. ALD by Grade, Test Event, and Strand – Grade 3

Achievement Level	Test Event	Strand	Number of Items	ALD 1	ALD 2	ALD 3
Developing	1	RP	8	1 (13%)	5 (63%)	2 (25%)
		RI	10	1 (10%)	7 (70%)	2 (20%)
		V	5	0 (0%)	4 (80%)	1 (20%)
		W	6	0 (0%)	6 (100%)	0 (0%)
	2	RP	7	3 (43%)	2 (29%)	2 (29%)
		RI	10	1 (10%)	8 (80%)	1 (10%)
		V	9	2 (22%)	5 (56%)	2 (22%)
		W	6	0 (0%)	6 (100%)	0 (0%)
	3	RP	7	0 (0%)	5 (71%)	2 (29%)
		RI	11	0 (0%)	6 (55%)	5 (45%)
		V	8	0 (0%)	6 (75%)	2 (25%)
		W	6	1 (17%)	5 (83%)	0 (0%)
	4	RP	8	1 (13%)	5 (63%)	2 (25%)
		RI	9	1 (11%)	5 (56%)	3 (33%)
		V	8	0 (0%)	7 (88%)	1 (13%)
		W	6	0 (0%)	6 (100%)	0 (0%)
On Track	1	RP	9	2 (22%)	3 (33%)	4 (44%)
		RI	6	0 (0%)	2 (33%)	4 (67%)
		V	6	0 (0%)	3 (50%)	3 (50%)
		W	6	1 (17%)	4 (67%)	1 (17%)
	2	RP	9	0 (0%)	7 (78%)	2 (22%)
		RI	7	2 (29%)	5 (71%)	0 (0%)
		V	5	0 (0%)	3 (60%)	2 (40%)
		W	6	1 (17%)	4 (67%)	1 (17%)
	3	RP	8	2 (25%)	4 (50%)	2 (25%)
		RI	8	0 (0%)	8 (100%)	0 (0%)
		V	8	0 (0%)	5 (63%)	3 (38%)
		W	6	1 (17%)	5 (83%)	0 (0%)
	4	RP	7	1 (14%)	5 (71%)	1 (14%)
		RI	8	0 (0%)	7 (88%)	1 (13%)
		V	5	0 (0%)	1 (20%)	4 (80%)
		W	6	0 (0%)	6 (100%)	0 (0%)



Achievement Level	Test Event	Strand	Number of Items	ALD 1	ALD 2	ALD 3
Advanced	1	RP	9	1 (11%)	6 (67%)	2 (22%)
		RI	12	0 (0%)	9 (75%)	3 (25%)
		V	4	0 (0%)	0 (0%)	4 (100%)
		W	6	0 (0%)	5 (83%)	1 (17%)
	2	RP	9	3 (33%)	3 (33%)	3 (33%)
		RI	8	0 (0%)	4 (50%)	4 (50%)
		V	5	0 (0%)	2 (40%)	3 (60%)
		W	6	0 (0%)	6 (100%)	0 (0%)
	3	RP	8	0 (0%)	6 (75%)	2 (25%)
		RI	7	1 (14%)	5 (71%)	1 (14%)
		V	7	0 (0%)	1 (14%)	6 (86%)
		W	6	0 (0%)	6 (100%)	0 (0%)
	4	RP	10	0 (0%)	8 (80%)	2 (20%)
		RI	5	1 (20%)	4 (80%)	0 (0%)
		V	6	0 (0%)	3 (50%)	3 (50%)
		W	6	1 (17%)	5 (83%)	0 (0%)



Table K2. ALD by Grade, Test Event, and Strand – Grade 4

Achievement Level	Test Event	Strand	Number of Items	ALD 1	ALD 2	ALD 3
Developing	1	RP	8	3 (38%)	4 (50%)	1 (13%)
		RI	8	5 (63%)	3 (38%)	0 (0%)
		V	6	1 (17%)	5 (83%)	0 (0%)
		W	6	0 (0%)	5 (83%)	1 (17%)
	2	RP	7	2 (29%)	5 (71%)	0 (0%)
		RI	4	4 (100%)	0 (0%)	0 (0%)
		V	6	1 (17%)	5 (83%)	0 (0%)
		W	6	1 (17%)	4 (67%)	1 (17%)
	3	RP	10	2 (20%)	7 (70%)	1 (10%)
		RI	6	3 (50%)	2 (33%)	1 (17%)
		V	7	1 (14%)	5 (71%)	1 (14%)
		W	6	2 (33%)	4 (67%)	0 (0%)
	4	RP	7	1 (14%)	5 (71%)	1 (14%)
		RI	6	4 (67%)	2 (33%)	0 (0%)
		V	5	1 (20%)	3 (60%)	1 (20%)
		W	6	1 (17%)	4 (67%)	1 (17%)
On Track	1	RP	7	3 (43%)	4 (57%)	0 (0%)
		RI	7	5 (71%)	2 (29%)	0 (0%)
		V	3	0 (0%)	2 (67%)	1 (33%)
		W	6	1 (17%)	4 (67%)	1 (17%)
	2	RP	9	3 (33%)	5 (56%)	1 (11%)
		RI	6	3 (50%)	3 (50%)	0 (0%)
		V	6	1 (17%)	4 (67%)	1 (17%)
		W	6	1 (17%)	4 (67%)	1 (17%)
	3	RP	7	2 (29%)	4 (57%)	1 (14%)
		RI	8	4 (50%)	4 (50%)	0 (0%)
		V	5	0 (0%)	1 (20%)	4 (80%)
		W	6	1 (17%)	5 (83%)	0 (0%)
	4	RP	9	3 (33%)	3 (33%)	3 (33%)
		RI	7	3 (43%)	4 (57%)	0 (0%)
		V	4	1 (25%)	2 (50%)	1 (25%)
		W	6	1 (17%)	5 (83%)	0 (0%)



Achievement Level	Test Event	Strand	Number of Items	ALD 1	ALD 2	ALD 3
Advanced	1	RP	8	2 (25%)	6 (75%)	0 (0%)
		RI	8	4 (50%)	4 (50%)	0 (0%)
		V	5	1 (20%)	2 (40%)	2 (40%)
		W	6	2 (33%)	4 (67%)	0 (0%)
	2	RP	7	1 (14%)	6 (86%)	0 (0%)
		RI	8	3 (38%)	5 (63%)	0 (0%)
		V	4	0 (0%)	3 (75%)	1 (25%)
		W	6	1 (17%)	4 (67%)	1 (17%)
	3	RP	7	1 (14%)	4 (57%)	2 (29%)
		RI	7	2 (29%)	5 (71%)	0 (0%)
		V	6	1 (17%)	3 (50%)	2 (33%)
		W	6	1 (17%)	5 (83%)	0 (0%)
	4	RP	8	3 (38%)	3 (38%)	2 (25%)
		RI	7	2 (29%)	4 (57%)	1 (14%)
		V	8	1 (13%)	6 (75%)	1 (13%)
		W	6	1 (17%)	5 (83%)	0 (0%)



Table K3. ALD by Grade, Test Event, and Strand – Grade 5

Achievement Level	Test Event	Strand	Number of Items	ALD 1	ALD 2	ALD 3
Developing	1	RP	10	1 (10%)	9 (90%)	0 (0%)
		RI	10	0 (0%)	10 (100%)	0 (0%)
		V	6	2 (33%)	4 (67%)	0 (0%)
		W	6	1 (17%)	5 (83%)	0 (0%)
	2	RP	9	1 (11%)	8 (89%)	0 (0%)
		RI	4	0 (0%)	4 (100%)	0 (0%)
		V	6	3 (50%)	3 (50%)	0 (0%)
		W	6	0 (0%)	6 (100%)	0 (0%)
	3	RP	7	2 (29%)	5 (71%)	0 (0%)
		RI	10	3 (30%)	6 (60%)	1 (10%)
		V	6	1 (17%)	5 (83%)	0 (0%)
		W	6	1 (17%)	5 (83%)	0 (0%)
	4	RP	11	3 (27%)	8 (73%)	0 (0%)
		RI	8	1 (13%)	7 (88%)	0 (0%)
		V	7	2 (29%)	5 (71%)	0 (0%)
		W	6	1 (17%)	5 (83%)	0 (0%)
On Track	1	RP	7	0 (0%)	7 (100%)	0 (0%)
		RI	8	0 (0%)	8 (100%)	0 (0%)
		V	5	4 (80%)	1 (20%)	0 (0%)
		W	6	0 (0%)	6 (100%)	0 (0%)
	2	RP	6	1 (17%)	5 (83%)	0 (0%)
		RI	9	1 (11%)	8 (89%)	0 (0%)
		V	7	2 (29%)	5 (71%)	0 (0%)
		W	6	2 (33%)	4 (67%)	0 (0%)
	3	RP	7	0 (0%)	7 (100%)	0 (0%)
		RI	8	0 (0%)	8 (100%)	0 (0%)
		V	9	2 (22%)	7 (78%)	0 (0%)
		W	6	0 (0%)	6 (100%)	0 (0%)
	4	RP	7	1 (14%)	6 (86%)	0 (0%)
		RI	6	2 (33%)	4 (67%)	0 (0%)
		V	2	0 (0%)	1 (50%)	1 (50%)
		W	6	1 (17%)	5 (83%)	0 (0%)



Achievement Level	Test Event	Strand	Number of Items	ALD 1	ALD 2	ALD 3
Advanced	1	RP	11	3 (27%)	8 (73%)	0 (0%)
		RI	7	0 (0%)	7 (100%)	0 (0%)
		V	5	3 (60%)	2 (40%)	0 (0%)
		W	6	2 (33%)	4 (67%)	0 (0%)
	2	RP	7	0 (0%)	7 (100%)	0 (0%)
		RI	10	1 (10%)	8 (80%)	1 (10%)
		V	7	1 (14%)	6 (86%)	0 (0%)
		W	6	2 (33%)	4 (67%)	0 (0%)
	3	RP	5	1 (20%)	4 (80%)	0 (0%)
		RI	7	3 (43%)	4 (57%)	0 (0%)
		V	4	0 (0%)	4 (100%)	0 (0%)
		W	6	2 (33%)	4 (67%)	0 (0%)
	4	RP	1	0 (0%)	1 (100%)	0 (0%)
		RI	10	0 (0%)	9 (90%)	1 (10%)
		V	7	2 (29%)	5 (71%)	0 (0%)
		W	6	2 (33%)	4 (67%)	0 (0%)



Table K4. ALD by Grade, Test Event, and Strand – Grade 6

Achievement Level	Test Event	Strand	Number of Items	ALD 1	ALD 2	ALD 3
Developing	1	RP	10	1 (10%)	9 (90%)	0 (0%)
		RI	3	0 (0%)	3 (100%)	0 (0%)
		V	4	1 (25%)	3 (75%)	0 (0%)
		W	6	1 (17%)	5 (83%)	0 (0%)
	2	RP	7	4 (57%)	3 (43%)	0 (0%)
		RI	9	4 (44%)	5 (56%)	0 (0%)
		V	5	2 (40%)	3 (60%)	0 (0%)
		W	6	0 (0%)	6 (100%)	0 (0%)
	3	RP	7	3 (43%)	3 (43%)	1 (14%)
		RI	9	3 (33%)	6 (67%)	0 (0%)
		V	7	2 (29%)	4 (57%)	1 (14%)
		W	6	0 (0%)	6 (100%)	0 (0%)
	4	RP	8	1 (13%)	6 (75%)	1 (13%)
		RI	9	3 (33%)	6 (67%)	0 (0%)
		V	6	2 (33%)	3 (50%)	1 (17%)
		W	6	0 (0%)	6 (100%)	0 (0%)
On Track	1	RP	7	1 (14%)	6 (86%)	0 (0%)
		RI	6	1 (17%)	4 (67%)	1 (17%)
		V	8	0 (0%)	6 (75%)	2 (25%)
		W	6	0 (0%)	6 (100%)	0 (0%)
	2	RP	10	1 (10%)	7 (70%)	2 (20%)
		RI	7	1 (14%)	6 (86%)	0 (0%)
		V	7	2 (29%)	5 (71%)	0 (0%)
		W	6	0 (0%)	6 (100%)	0 (0%)
	3	RP	8	0 (0%)	7 (88%)	1 (13%)
		RI	6	2 (33%)	4 (67%)	0 (0%)
		V	7	0 (0%)	7 (100%)	0 (0%)
		W	6	0 (0%)	6 (100%)	0 (0%)
	4	RP	8	0 (0%)	5 (63%)	3 (38%)
		RI	8	1 (13%)	6 (75%)	1 (13%)
		V	6	0 (0%)	4 (67%)	2 (33%)
		W	6	0 (0%)	6 (100%)	0 (0%)



Achievement Level	Test Event	Strand	Number of Items	ALD 1	ALD 2	ALD 3
Advanced	1	RP	7	1 (14%)	5 (71%)	1 (14%)
		RI	9	2 (22%)	7 (78%)	0 (0%)
		V	7	1 (14%)	6 (86%)	0 (0%)
		W	6	1 (17%)	5 (83%)	0 (0%)
	2	RP	7	1 (14%)	4 (57%)	2 (29%)
		RI	9	1 (11%)	7 (78%)	1 (11%)
		V	7	1 (14%)	5 (71%)	1 (14%)
		W	6	1 (17%)	5 (83%)	0 (0%)
	3	RP	8	0 (0%)	7 (88%)	1 (13%)
		RI	7	0 (0%)	7 (100%)	0 (0%)
		V	5	0 (0%)	5 (100%)	0 (0%)
		W	6	1 (17%)	5 (83%)	0 (0%)
	4	RP	8	0 (0%)	6 (75%)	2 (25%)
		RI	9	3 (33%)	6 (67%)	0 (0%)
		V	5	2 (40%)	3 (60%)	0 (0%)
		W	6	0 (0%)	6 (100%)	0 (0%)



Table K5. ALD by Grade, Test Event, and Strand – Grade 7

Achievement Level	Test Event	Strand	Number of Items	ALD 1	ALD 2	ALD 3
Developing	1	RP	7	1 (14%)	5 (71%)	1 (14%)
		RI	11	7 (64%)	4 (36%)	0 (0%)
		V	7	2 (29%)	2 (29%)	3 (43%)
		W	5	2 (40%)	3 (60%)	0 (0%)
	2	RP	7	1 (14%)	5 (71%)	1 (14%)
		RI	9	2 (22%)	7 (78%)	0 (0%)
		V	7	0 (0%)	5 (71%)	2 (29%)
		W	6	1 (17%)	5 (83%)	0 (0%)
	3	RP	7	1 (14%)	4 (57%)	2 (29%)
		RI	9	3 (33%)	5 (56%)	0 (0%)
		V	6	0 (0%)	2 (33%)	4 (67%)
		W	5	0 (0%)	4 (80%)	1 (20%)
	4	RP	8	2 (25%)	4 (50%)	2 (25%)
		RI	8	2 (25%)	6 (75%)	0 (0%)
		V	9	2 (22%)	2 (22%)	5 (56%)
		W	5	2 (40%)	3 (60%)	0 (0%)
On Track	1	RP	4	0 (0%)	2 (50%)	2 (50%)
		RI	9	3 (33%)	4 (44%)	2 (22%)
		V	4	0 (0%)	1 (25%)	3 (75%)
		W	6	1 (17%)	5 (83%)	0 (0%)
	2	RP	9	1 (11%)	7 (78%)	1 (11%)
		RI	7	5 (71%)	1 (14%)	1 (14%)
		V	4	0 (0%)	3 (75%)	1 (25%)
		W	6	0 (0%)	6 (100%)	0 (0%)
	3	RP	7	1 (14%)	5 (71%)	1 (14%)
		RI	10	3 (30%)	7 (70%)	0 (0%)
		V	5	0 (0%)	4 (80%)	1 (20%)
		W	6	2 (33%)	3 (50%)	1 (17%)
	4	RP	8	2 (25%)	5 (63%)	1 (13%)
		RI	8	2 (25%)	5 (63%)	1 (13%)
		V	5	1 (20%)	3 (60%)	1 (20%)
		W	6	1 (17%)	5 (83%)	0 (0%)



Achievement Level	Test Event	Strand	Number of Items	ALD 1	ALD 2	ALD 3
Advanced	1	RP	7	2 (29%)	4 (57%)	1 (14%)
		RI	9	3 (33%)	5 (56%)	1 (11%)
		V	5	0 (0%)	3 (60%)	2 (40%)
		W	6	0 (0%)	4 (67%)	2 (33%)
	2	RP	5	0 (0%)	4 (80%)	1 (20%)
		RI	7	1 (14%)	4 (57%)	2 (29%)
		V	5	0 (0%)	2 (40%)	3 (60%)
		W	4	0 (0%)	4 (100%)	0 (0%)
	3	RP	6	0 (0%)	3 (50%)	3 (50%)
		RI	9	2 (22%)	7 (78%)	0 (0%)
		V	5	0 (0%)	4 (80%)	1 (20%)
		W	6	1 (17%)	4 (67%)	1 (17%)
	4	RP	7	2 (29%)	5 (71%)	0 (0%)
		RI	9	1 (11%)	7 (78%)	1 (11%)
		V	5	0 (0%)	4 (80%)	1 (20%)
		W	5	0 (0%)	4 (80%)	1 (20%)



Table K6. ALD by Grade, Test Event, and Strand – Grade 8

Achievement Level	Test Event	Strand	Number of Items	ALD 1	ALD 2	ALD 3
Developing	1	RP	9	0 (0%)	8 (89%)	1 (11%)
		RI	6	1 (17%)	5 (83%)	0 (0%)
		V	7	2 (29%)	4 (57%)	1 (14%)
		W	6	1 (17%)	5 (83%)	0 (0%)
	2	RP	11	3 (27%)	7 (64%)	1 (9%)
		RI	9	4 (44%)	5 (56%)	0 (0%)
		V	4	1 (25%)	3 (75%)	0 (0%)
		W	5	2 (40%)	3 (60%)	0 (0%)
	3	RP	4	1 (25%)	2 (50%)	1 (25%)
		RI	5	2 (40%)	3 (60%)	0 (0%)
		V	6	1 (17%)	5 (83%)	0 (0%)
		W	6	0 (0%)	5 (83%)	1 (17%)
	4	RP	7	1 (14%)	5 (71%)	1 (14%)
		RI	9	4 (44%)	5 (56%)	0 (0%)
		V	10	1 (10%)	9 (90%)	0 (0%)
		W	6	1 (17%)	4 (67%)	1 (17%)
On Track	1	RP	9	1 (11%)	8 (89%)	0 (0%)
		RI	7	0 (0%)	7 (100%)	0 (0%)
		V	5	1 (20%)	4 (80%)	0 (0%)
		W	6	0 (0%)	6 (100%)	0 (0%)
	2	RP	7	0 (0%)	4 (57%)	3 (43%)
		RI	3	2 (67%)	1 (33%)	0 (0%)
		V	5	0 (0%)	5 (100%)	0 (0%)
		W	6	0 (0%)	6 (100%)	0 (0%)
	3	RP	7	1 (14%)	6 (86%)	0 (0%)
		RI	9	2 (22%)	5 (56%)	2 (22%)
		V	7	0 (0%)	7 (100%)	0 (0%)
		W	6	0 (0%)	6 (100%)	0 (0%)
	4	RP	7	1 (14%)	6 (86%)	0 (0%)
		RI	8	3 (38%)	4 (50%)	1 (13%)
		V	8	0 (0%)	7 (88%)	1 (13%)
		W	6	1 (17%)	5 (83%)	0 (0%)



Achievement Level	Test Event	Strand	Number of Items	ALD 1	ALD 2	ALD 3
Advanced	1	RP	8	0 (0%)	6 (75%)	2 (25%)
		RI	7	1 (14%)	5 (71%)	1 (14%)
		V	6	0 (0%)	5 (83%)	1 (17%)
		W	6	0 (0%)	6 (100%)	0 (0%)
	2	RP	7	1 (14%)	6 (86%)	0 (0%)
		RI	9	0 (0%)	6 (67%)	3 (33%)
		V	5	0 (0%)	5 (100%)	0 (0%)
		W	6	0 (0%)	6 (100%)	0 (0%)
	3	RP	4	0 (0%)	1 (25%)	3 (75%)
		RI	9	1 (11%)	8 (89%)	0 (0%)
		V	6	1 (17%)	3 (50%)	2 (33%)
		W	6	0 (0%)	6 (100%)	0 (0%)
	4	RP	8	2 (25%)	4 (50%)	2 (25%)
		RI	7	3 (43%)	3 (43%)	1 (14%)
		V	4	0 (0%)	4 (100%)	0 (0%)
		W	6	0 (0%)	6 (100%)	0 (0%)



Appendix L. Process Evaluation Tables by Grade

Table L1. Panelist Evaluation Survey Results – Grade 3

Question	Average	
My panel facilitator did an effective job of facilitating discussion and ensuring that all panelists' perspectives were heard	4.75	<div></div>
The panel facilitator was helpful during the workshop	4.75	<div></div>
Everyone had equal opportunity to contribute ideas and opinions	4.50	<div></div>
My ideas and opinions were listened to and respected by the group	4.50	<div></div>
My panel facilitator clearly and promptly addressed my questions	4.50	<div></div>
The other materials shared by my facilitator were useful	4.50	<div></div>
The whole-group training facilitator was helpful during the workshop	4.50	<div></div>
It was easy to access the evaluation and demographics forms	4.25	<div></div>
It was easy to access the items on the Content Review Tool	4.25	<div></div>
Other support staff were helpful during the workshop	4.25	<div></div>
Practicing making ratings as a group in my assigned panel helped me better understand the alignment activities	4.25	<div></div>
The Content Review Tool allowed me to effectively accomplish my tasks during the workshop	4.25	<div></div>
The Google Rating Sheet was useful for recording alignment ratings	4.25	<div></div>
The hands-on training in my assigned panel helped me better understand the alignment activities	4.25	<div></div>
The hands-on training in my assigned panel was an effective use of time	4.25	<div></div>
The materials hosted on Google Drive were useful (e.g., standards)	4.25	<div></div>
The panel-specific hands-on training was well organized	4.25	<div></div>
The Google Rating Sheet provided a comprehensive platform for capturing alignment of standards	4.00	<div></div>
The group-wide training session provided a useful overview of the alignment activities for the week	4.00	<div></div>
The group-wide training session was well organized	4.00	<div></div>
The group-wide training was an effective use of time	4.00	<div></div>
The group-wide training session clearly described my role as a panelist	3.75	<div></div>
The group-wide training session effectively outlined the purpose of the alignment workshop	3.75	<div></div>

Note: Strongly Disagree = 1 to Strongly Agree = 5



Table L2. Panelist Evaluation Survey Results – Grade 4

Question	Average	
Practicing making ratings as a group in my assigned panel helped me better understand the alignment activities	4.67	<div></div>
The hands-on training in my assigned panel was an effective use of time	4.67	<div></div>
My panel facilitator clearly and promptly addressed my questions	4.67	<div></div>
Everyone had equal opportunity to contribute ideas and opinions	4.67	<div></div>
My ideas and opinions were listened to and respected by the group	4.67	<div></div>
The hands-on training in my assigned panel helped me better understand the alignment activities	4.33	<div></div>
The panel-specific hands-on training was well organized	4.33	<div></div>
My panel facilitator did an effective job of facilitating discussion and ensuring that all panelists' perspectives were heard	4.33	<div></div>
The panel facilitator was helpful during the workshop	4.33	<div></div>
The group-wide training session effectively outlined the purpose of the alignment workshop	4.00	<div></div>
The group-wide training session provided a useful overview of the alignment activities for the week	4.00	<div></div>
The group-wide training session clearly described my role as a panelist	4.00	<div></div>
The group-wide training session was well organized	4.00	<div></div>
The group-wide training was an effective use of time	4.00	<div></div>
The materials hosted on Google Drive were useful (e.g., standards)	4.00	<div></div>
The Google Rating Sheet provided a comprehensive platform for capturing alignment of standards	4.00	<div></div>
It was easy to access the items on the Content Review Tool	4.00	<div></div>
The Content Review Tool allowed me to effectively accomplish my tasks during the workshop	4.00	<div></div>
It was easy to access the evaluation and demographics forms	4.00	<div></div>
The whole-group training facilitator was helpful during the workshop	4.00	<div></div>
The other materials shared by my facilitator were useful	3.67	<div></div>
Other support staff were helpful during the workshop	3.67	<div></div>
The Google Rating Sheet was useful for recording alignment ratings	3.33	<div></div>

Note: Strongly Disagree = 1 to Strongly Agree = 5



Table L3. Panelist Evaluation Survey Results – Grade 5

Question	Average	
My panel facilitator clearly and promptly addressed my questions	5.00	
My panel facilitator did an effective job of facilitating discussion and ensuring that all panelists' perspectives were heard	5.00	
The panel facilitator was helpful during the workshop	5.00	
The materials hosted on Google Drive were useful (e.g., standards)	4.80	
It was easy to access the evaluation and demographics forms	4.80	
Practicing making ratings as a group in my assigned panel helped me better understand the alignment activities	4.60	
The Google Rating Sheet was useful for recording alignment ratings	4.60	
The Google Rating Sheet provided a comprehensive platform for capturing alignment of standards	4.60	
The other materials shared by my facilitator were useful	4.60	
The group-wide training session was well organized	4.50	
The group-wide training session effectively outlined the purpose of the alignment workshop	4.40	
The group-wide training was an effective use of time	4.40	
The panel-specific hands-on training was well organized	4.40	
The hands-on training in my assigned panel was an effective use of time	4.40	
Everyone had equal opportunity to contribute ideas and opinions	4.40	
My ideas and opinions were listened to and respected by the group	4.40	
The whole-group training facilitator was helpful during the workshop	4.40	
Other support staff were helpful during the workshop	4.40	
The group-wide training session provided a useful overview of the alignment activities for the week	4.20	
The group-wide training session clearly described my role as a panelist	4.20	
The hands-on training in my assigned panel helped me better understand the alignment activities	4.20	
It was easy to access the items on the Content Review Tool	4.00	
The Content Review Tool allowed me to effectively accomplish my tasks during the workshop	4.00	

Note: Strongly Disagree = 1 to Strongly Agree = 5



Table L4. Panelist Evaluation Survey Results – Grade 6

Question	Average	
My panel facilitator clearly and promptly addressed my questions	5.00	<div></div>
My panel facilitator did an effective job of facilitating discussion and ensuring that all panelists' perspectives were heard	5.00	<div></div>
The materials hosted on Google Drive were useful (e.g., standards)	5.00	<div></div>
The Google Rating Sheet was useful for recording alignment ratings	5.00	<div></div>
The hands-on training in my assigned panel helped me better understand the alignment activities	4.83	<div></div>
Practicing making ratings as a group in my assigned panel helped me better understand the alignment activities	4.83	<div></div>
The hands-on training in my assigned panel was an effective use of time	4.83	<div></div>
Everyone had equal opportunity to contribute ideas and opinions	4.83	<div></div>
My ideas and opinions were listened to and respected by the group	4.83	<div></div>
The Google Rating Sheet provided a comprehensive platform for capturing alignment of standards	4.83	<div></div>
The other materials shared by my facilitator were useful	4.83	<div></div>
The whole-group training facilitator was helpful during the workshop	4.83	<div></div>
The panel facilitator was helpful during the workshop	4.83	<div></div>
Other support staff were helpful during the workshop	4.83	<div></div>
The panel-specific hands-on training was well organized	4.67	<div></div>
The Content Review Tool allowed me to effectively accomplish my tasks during the workshop	4.50	<div></div>
It was easy to access the evaluation and demographics forms	4.50	<div></div>
The group-wide training session effectively outlined the purpose of the alignment workshop	4.17	<div></div>
The group-wide training session provided a useful overview of the alignment activities for the week	4.17	<div></div>
The group-wide training session clearly described my role as a panelist	4.17	<div></div>
The group-wide training session was well organized	4.17	<div></div>
It was easy to access the items on the Content Review Tool	4.17	<div></div>
The group-wide training was an effective use of time	4.00	<div></div>

Note: Strongly Disagree = 1 to Strongly Agree = 5



Table L5. Panelist Evaluation Survey Results – Grade 7

Question	Average	
My panel facilitator clearly and promptly addressed my questions	5.00	
My panel facilitator did an effective job of facilitating discussion and ensuring that all panelists' perspectives were heard	5.00	
Everyone had equal opportunity to contribute ideas and opinions	5.00	
My ideas and opinions were listened to and respected by the group	5.00	
The group-wide training session effectively outlined the purpose of the alignment workshop	4.80	
The group-wide training session provided a useful overview of the alignment activities for the week	4.80	
The group-wide training session clearly described my role as a panelist	4.80	
The group-wide training session was well organized	4.80	
The group-wide training was an effective use of time	4.80	
Practicing making ratings as a group in my assigned panel helped me better understand the alignment activities	4.80	
The panel-specific hands-on training was well organized	4.80	
The hands-on training in my assigned panel was an effective use of time	4.80	
The materials hosted on Google Drive were useful (e.g., standards)	4.80	
The Google Rating Sheet was useful for recording alignment ratings	4.80	
The Google Rating Sheet provided a comprehensive platform for capturing alignment of standards	4.80	
The other materials shared by my facilitator were useful	4.80	
It was easy to access the evaluation and demographics forms	4.80	
The whole-group training facilitator was helpful during the workshop	4.80	
The panel facilitator was helpful during the workshop	4.80	
Other support staff were helpful during the workshop	4.80	
The hands-on training in my assigned panel helped me better understand the alignment activities	4.60	
The Content Review Tool allowed me to effectively accomplish my tasks during the workshop	4.60	
It was easy to access the items on the Content Review Tool	4.20	

Note: Strongly Disagree = 1 to Strongly Agree = 5



Table L6. Panelist Evaluation Survey Results – Grade 8

Question	Average	
Practicing making ratings as a group in my assigned panel helped me better understand the alignment activities	5.00	
The panel-specific hands-on training was well organized	5.00	
The hands-on training in my assigned panel was an effective use of time	5.00	
My panel facilitator clearly and promptly addressed my questions	5.00	
My panel facilitator did an effective job of facilitating discussion and ensuring that all panelists' perspectives were heard	5.00	
Everyone had equal opportunity to contribute ideas and opinions	5.00	
My ideas and opinions were listened to and respected by the group	5.00	
The materials hosted on Google Drive were useful (e.g., standards)	5.00	
The panel facilitator was helpful during the workshop	5.00	
The group-wide training session clearly described my role as a panelist	4.86	
The Google Rating Sheet was useful for recording alignment ratings	4.86	
The Google Rating Sheet provided a comprehensive platform for capturing alignment of standards	4.86	
The other materials shared by my facilitator were useful	4.86	
It was easy to access the evaluation and demographics forms	4.86	
The group-wide training session effectively outlined the purpose of the alignment workshop	4.71	
The group-wide training session provided a useful overview of the alignment activities for the week	4.71	
The group-wide training session was well organized	4.71	
The group-wide training was an effective use of time	4.57	
The hands-on training in my assigned panel helped me better understand the alignment activities	4.57	
The Content Review Tool allowed me to effectively accomplish my tasks during the workshop	4.57	
The whole-group training facilitator was helpful during the workshop	4.57	
Other support staff were helpful during the workshop	4.43	
It was easy to access the items on the Content Review Tool	4.14	

Note: Strongly Disagree = 1 to Strongly Agree = 5

**Table L7. Overall Alignment – All grades**

Answers	Count	Percentage
Strongly aligned	15	50%
Partially aligned	15	50%
Not at all aligned	0	0%

Table L8. Overall Alignment – Grade 3

Answers	Count	Percentage
Strongly aligned	2	50%
Partially aligned	2	50%
Not at all aligned	0	0%

Table L9. Overall Alignment – Grade 4

Answers	Count	Percentage
Strongly aligned	3	100%
Partially aligned	0	0%
Not at all aligned	0	0%

Table L10. Overall Alignment – Grade 5

Answers	Count	Percentage
Strongly aligned	4	80%
Partially aligned	1	20%
Not at all aligned	0	0%

Table L11. Overall Alignment – Grade 6

Answers	Count	Percentage
Strongly aligned	1	16.7%
Partially aligned	5	83.3%
Not at all aligned	0	0%

Table L12. Overall Alignment – Grade 7

Answers	Count	Percentage
Strongly aligned	1	20%
Partially aligned	4	80%
Not at all aligned	0	0%

Table L13. Overall Alignment – Grade 8

Answers	Count	Percentage
Strongly aligned	4	57.1%
Partially aligned	3	42.9%
Not at all aligned	0	0%